Die Validität der Interpretationen studentischer Lehrevaluationsergebnisse:

Eine exemplarische Anwendung des argumentationsbasierten Ansatzes

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie

in Frankfurt am Main

Vorgelegt beim

Fachbereich Psychologie und Sportwissenschaften

der Johann Wolfgang Goethe - Universität

Frankfurt am Main

Vorgelegt von

Michael Anton Paulitsch

aus Offenbach am Main

Frankfurt am Main, 2017

(D.30)

Dekan:	Prof. Dr. Rolf van Dick
Gutachter:	Prof. Dr. Johannes Hartig
	Prof. Dr. Holger Horz

„Validation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle combining the attractions of chess and mud wrestling." Lee J. Cronbach (1988)

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

IV. Abkürzungsverzeichnis

AERA	American Educational Research Association
APA	American Psychological Association
AUA	Assessment Use Argument
ECD	Evidence Centered Design
EFA	Exploratorische Faktorenanalyse
FESEM	Fragebogen zur Evaluation von Seminaren
FEVOR	Fragebogen zur Evaluation von Vorlesungen
GT	Generalisierbarkeitstheorie
HILVE	Heidelberger Inventar zur Lehrveranstaltungsevaluation
IA	Interpretative Argument
IRT	Item Response - Theorie
IUA	Interpretation/Use - Argument
KTT	Klassische Testtheorie
NCME	National Council on Measurement in Education
NSSE	National Survey of Student Engagement
PK	Promotionskolleg

Zusammenfassung

Studentische Lehrevaluationsergebnisse sind ein weit verbreitetes Maß, um die Qualität universitärer Lehre zu erfassen. Diese Ergebnisse werden unter anderem dafür genutzt, Entscheidungen für die Modifikation des Lehrangebots zu treffen oder die Vergabe der Leistungsorientieren Mittelvergabe mitzubestimmen. Aufgrund dieser relevanten Folgen wird in dieser Arbeit der Frage nachgegangen, wie ein angemessener Validierungsprozess bezüglich studentischer Lehrevaluationsergebnisse gestaltet werden könnte.

Bisherige Validierungsstudien zu studentischen Lehrevaluationsinventaren fokussierten sich meist auf die Überprüfung verschiedener Validitätsarten (inhaltsbezogene, kriteriumsbezogene oder faktorielle) und die Erfassung der Messfehlerfreiheit.

Allerdings ist zum einen zu hinterfragen, ob diese Ansätze grundsätzlich für alle Inventare geeignet sind. Weiterhin hat sich das Verständnis von dem verändert, was unter Validität verstanden wird: Von der Annahme von Validität als Testeigenschaft, verschiedener Validitätsarten und binärer Aussagen auf Basis von Einzelbefunden hin zu dem Verständnis von Validität bezogen auf die Testwert-Interpretation und Verwendung, zu einem einheitlichen Validitätskonzept und zu einer Validitäts-Argumentation. Diese Veränderungen werden in den neueren argumentationsbasierten Validitätsansätzen berücksichtigt und bieten einen Rahmen, der auf die jeweilige Intention ausgerichtet ist, einen Test oder Fragebogen einzusetzen.

Auf Grundlage dieser argumentationsbasierten Ansätze wird in dieser Arbeit die Interpretation studentischer Lehrevaluationsergebnisse überprüft, die als das Ausmaß an qualitätsbezogener Zufriedenheit der Teilnehmer mit der Durchführung einer Lehrveranstaltung und der Vermittlung von Lehrinhalten angesehen werden. Der Validierungsprozess wird anhand der Lehrevaluationsdaten des Frankfurter Promotionskollegs am Fachbereich Medizin dargestellt. Dieser Prozess bestätigte weitgehend die beabsichtigte Interpretation, zeigte aber auch eine zumindest teilweise Revision des Inventars und eine weitere Überprüfung an. Eine Validierung bezüglich der Verwendung der Lehrevaluationsergebnisse sowie der auf diesen basierenden beabsichtigten Konsequenzen wird in einer Folgestudie überprüft.

Anhand dieser Arbeit wird Anwendern und Entwicklern von Lehrevaluationsinventaren eine Her- und Anleitung für den Validierungsprozess gegeben und die Vorteile argumentationsbasierter Ansätze aufgezeigt.

1. Hintergrund

Die Validitätsbeurteilung studentischer Lehrevaluationsergebnisse im Kontext universitärer Lehre ist ein Forschungsgebiet mit praktischer Relevanz, wie folgende Aspekte veranschaulichen:

Zum einen ist die Evaluation von Lehre beziehungsweise Lehrveranstaltungen auf Basis studentischer Aussagen ein weit verbreiteter Bestandteil der Qualitätsbeurteilung von Universitäten. Zum Beispiel sind sie in nordamerikanischen Universitäten nahezu universell vorhanden (siehe Seldin, 1993; Centra, 2003) und breiteten sich auch in anderen Teilen der Welt aus (zum Beispiel in Australien, siehe Marsh & Roche, 1992). Auch in Deutschland zeigte sich eine breite Anwendung (Übersichten bei Souvignier & Gold, 2002; Braun, 2007; Rindermann, 2009). Es ist davon auszugehen, dass neben diesen Publikationen, die eine Übersicht zu Ergebnissen und Validierungsstudien publizierter Lehrevaluationsinventare liefern, weitere national als auch international konstruierte Inventare nicht in wissenschaftlichen Fachzeitschriften veröffentlicht wurden. Somit finden studentische Lehrevaluationen wahrscheinlich eine noch breitere Anwendung als eine Recherche in wissenschaftlichen Datenbanken nachweisen kann.

Weiterhin können die Ergebnisse studentischer Lehrevaluationen Entscheidungen beeinflussen oder sogar maßgeblich bestimmen, denen bedeutsame praktische Konsequenzen folgen: Beispielsweise werden Dozenten aufgrund nicht ausreichend positiver Beurteilung nicht mehr eingeladen, Lehrveranstaltungen abgesetzt oder deren Inhalte beziehungsweise die Art ihrer Vermittlung verändert. Zudem können sie neben Forschungsleistungen auch als Qualifikationsmaß bei Bleibe-, Gehalts- und Berufungsverhandlungen (Rindermann, 2009, S. 31) sowie zur *Leistungsorientierten Mittelvergabe* herangezogen werden.

Laut Rindermann (Rindermann, 2009, S. 31) werde neben der Rückmeldung zur Lehrverbesserung auch die Kommunikation zwischen Lehrenden und Studierenden über die Lehre gefördert und eine Informationsgrundlage geschaffen: Studierenden könne auf Basis der Ergebnisse eine Hilfestellung bei der Kurswahl gegeben werden und gemeinsam mit Berichten über Forschungsleistungen böten sie eine Grundlage, um Vergleiche zwischen einzelnen Universitäten vornehmen zu können.

Weiterhin sehen insbesondere sich entwickelnde Länder eine qualitativ höhere Bildung als notwendig für ihre ökonomische Entwicklung an (Altbach & Selvaratnam, 1989). Dementsprechend wichtig sind Instrumente zur Qualitätssicherung und Verbesserung vorhandener Angebote.

Neben diesen Aspekten fordert zusätzlich auch der gesellschaftspolitische Rahmen den Einbezug Studierender in die Lehrevaluation: Allgemein dienen Hochschulen nach § 3 des Hessischen Hochschulgesetzes „der Verwirklichung des Rechts auf Bildung durch Forschung, künstlerisches Schaffen, Lehre, Studium und Weiterbildung in einem freiheitlichen, demokratischen und sozialen Rechtsstaat“. Weiterhin bereitet der Besuch von Hochschulen „auf berufliche Aufgaben vor, bei denen die Anwendung wissenschaftlicher Erkenntnisse und Methoden oder die Fähigkeit zur künstlerischen Gestaltung erforderlich oder nützlich ist“. Dementsprechend sollen nach § 13, „Lehre und Studium […] wissenschaftlich-kritisches Denken und in entsprechenden Studiengängen künstlerische Fähigkeiten mit fachübergreifenden Bezügen“ vermitteln. Studierende sollen auf ein berufliches Tätigkeitsfeld vorbereitet werden und entsprechende fachlichen Kenntnisse und Methoden vermittelt bekommen. (Land Hessen, 2009)

Diese Vermittlung durch die universitäre Lehre zeigt sich in verschiedenen Facetten des universitären Alltags: Es werden Lehrveranstaltungen verschiedener Art durchgeführt, Abschlussarbeiten begleitet und Wissen in Bibliotheken erhalten und zur Vermittlung angeboten. Die Lehre an den jeweiligen Universitäten beziehungsweise Fakultäten wird im weiteren Sinne unter anderem auch durch die jeweiligen Studienzeiten und durch unterschiedliche Absolventenchancen charakterisiert. Letztendlich bilden Lehrveranstaltungen in Form von Vorlesungen, Seminaren und Tutorien den Schwerpunkt der Wissens- und Kompetenzvermittlung sowie der Kommunikation zwischen Lehrenden und Studierenden. Aufgrund dessen nimmt die *Lehrveranstaltungsevaluation* eine bedeutsame Stellung innerhalb der allgemeinen Lehrevaluation ein. (Rindermann, 2009, S. 27).

Im Zuge des Bologna-Prozesses (Die Europäischen Bildungsminister, 1999) wird der einzelnen Lehrveranstaltung mehr Bedeutung beigemessen: In den modularisierten Studiengängen zählen die Ergebnisse der einzelnen besuchten Lehrveranstaltung für die Studiumsabschlussnote. Aus dieser leitet sich auch bei Überbelegung die

Möglichkeit ab, einen weiterführenden Masterstudiengang zu besuchen. (zum Beispiel Fachbereich Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität Frankfurt am Main, 2013, 2014)

Eine Kontrolle, inwieweit Hochschulen die genannten Zwecke erfüllen, ist von Gesetzgebern empfohlen und gesetzlich vorgeschrieben: In § 6 des Hochschulrahmengesetzes von 1998 wurde dazu aufgefordert, eine Evaluation der Lehre durchzuführen (Bundesregierung der Bundesrepublik Deutschland, 1998). Auf Basis der *Qualitätssicherung und des Berichtswesens* nach § 12 des Hessischen Hochschulgesetzes evaluieren Hochschulen „regelmäßig die Erfüllung ihrer Aufgaben, insbesondere in Lehre, Forschung, Internationalisierung und interkultureller Integration, Wissens- und Technologietransfer, Weiterbildung und Verwaltung unter Berücksichtigung der Entwicklungen in Wissenschaft, Kunst, Gesellschaft und Berufswelt; hierbei sind in regelmäßigen Abständen externe Sachverständige hinzuzuziehen. An der Evaluation der Lehre sind die Studierenden zu beteiligen.“ [Hervorhebungen durch den Verfasser] (Land Hessen, 2009)

Aufgrund all dieser Anforderungen sowie tatsächlicher oder potenzieller Konsequenzen ist es für Entscheidungsträger in der universitären Lehre oder auch für die Lehrenden selbst wichtig, sich neben anderen Quellen (wie die Rückmeldung von Hochschuldidaktikern) auch auf die Ergebnisse studentischer Lehrevaluationen verlassen zu können - genauer ausgedrückt auf die ihnen zugewiesenen Bedeutung beziehungsweise Interpretation dieser Ergebnisse. Im wissenschaftlichen Kontext wird bei dieser grundsätzlichen Anforderung an Fragebögen oder Tests der Begriff der Validität verwendet. Was Validität ausmacht und wie sie nachgewiesen werden kann, ist seit vielen Jahrzehnten Bestandteil wissenschaftlicher Beiträge und Debatten (zum Beispiel Cronbach, 1988; Messick, 1989a; Markus & Borsboom, 2013; Kane, 2013).

Auch bezüglich der Erfüllung und des Nachweises wissenschaftlicher Kriterien bei der Erhebung studentischer Lehrevaluationsergebnisse wurde in den letzten Jahrzehnten eine schwer zu überblickende Fülle wissenschaftlicher Beiträge verfasst (nationale wie internationale Übersichten bei Souvignier & Gold, 2002; Braun, 2007; Marsh, 2007; Rindermann, 2009; Sippel, 2014).

Um also den genannten Zielen akademischer Bildung und Ausbildung gerecht zu werden und auch gesetzliche Vorgaben und Empfehlungen zu erfüllen, muss die Qualität dieser Lehrangebote gegeben sein. Dementsprechend sollte diese gemessen beziehungsweise überprüft werden. Um aus den Interpretationen der daraus resultierenden Ergebnisse angemessene Schlussfolgerungen beziehungsweise Konsequenzen ableiten zu können, muss vor dem Hintergrund der zu Beginn dieses Kapitels dargestellten Relevanz studentischer Lehrevaluationsergebnisse ein Validitätsnachweis dieser Interpretationen gegeben sein.

Vor diesem Hintergrund und auf den herkömmlichen Umgang zur Frage der Validität im Kontext studentischer Lehrevaluation aufbauend, wird in dieser Promotionsarbeit folgende Fragestellung untersucht:

> *Wie sollte ein Validierungsprozess gestaltet sein, in dem überprüft wird, ob studentische Lehrevaluationsergebnisse das beabsichtigte Konstrukt abbilden und für Entscheidungen mit entsprechenden Konsequenzen verwandt werden können?*

Um diese Fragestellung zu beantworten, werden folgende Aspekte untersucht und miteinander in Beziehung gesetzt:

1. Klärung, was durch ein studentisches Lehrevaluationsinventar erfasst werden soll.
2. Darstellung, in welcher Weise Validierungsstudien zu studentischen Lehrevaluationsinventaren in der Regel durchgeführt wurden.
3. Eine kritische Reflektion bisheriger Vorgehensweisen und Vorstellung von Alternativen.

Auf Basis der Schlussfolgerungen zu dieser Fragestellung wird anhand eines bereits seit 2011 eingesetzten studentischen Lehrevaluationsinventars am *Frankfurter Promotionskollegs des Fachbereichs Medizin* (Sennekamp, Paulitsch, Broermann, Klingebiel & Gerlach, 2016; Paulitsch, Gerlach, Klingebiel & Sennekamp, 2016) ein entsprechender Validierungsprozess anhand erhobener Daten exemplarisch beschrieben und angewandt.

Die folgende theoretische Diskussion und die empirische Veranschaulichung soll Entwicklern und Anwendern studentischer Lehrevaluationsinventare die

verschiedenen vorhandenen testtheoretischen[1] Ansätze der Validität aufzeigen sowie deren Vorzüge und Schwächen veranschaulichen. Schlussendlich soll mit dieser Arbeit eine Grundlage geschaffen werden, auf der Inventare nach wissenschaftlichen Kriterien gestaltet und einer Validitätsbeurteilung unterzogen werden können. Anwender solcher Inventare soll es ermöglicht werden, auf Basis eines wissenschaftlich fundierten Wissens, das für sie angemessene Inventar aus dem bereits vorhandenen Spektrum auswählen zu können.

[1] In dem Begriff des *Tests* werden in dieser Arbeit auch alle Formen von Fragebögen oder anderen Inventaren eingeschlossen.

2. Was macht Qualität der Lehre aus?

Qualitätsbeurteilungen werden in der Regel auch Evaluationen genannt. Vor der Durchführung einer Evaluation ist generell zu klären, was konkret beurteilt werden soll. In der vorliegenden Arbeit ist dies das Konzept der *Lehrqualität*.

Um universitäre Lehre angemessen evaluieren und somit den Grad ihrer Qualität bestimmen zu können, muss zunächst definiert werden, was in dem bestimmten Kontext unter Qualität verstanden wird und ab wann von ihrer Erfüllung ausgegangen werden kann.

Zunächst soll hier der Begriff der *Qualität* an sich geklärt werden. Auf Basis dieser Begriffserklärung wird dann eine für den Kontext dieser Arbeit angemessene Definition von Lehrqualität im Kontext studentischer Lehrevaluationen aufgestellt.

2.1 Was ist Qualität?

Diana Green (1994) stellt in dem von ihr herausgegebenen Buch „What is Quality in Higher Education“ dar, dass sich die Definitionen von *Qualität in der Höheren Bildung* je nach Perspektive und Auftrag unterscheiden: Gehe es um die Qualität der Zufuhr menschlicher und materieller Ressourcen, um die Qualität der Abschlüsse oder um die Qualität des Lehr- und Lernprozesses an sich?

Die Antwort unterscheide sich jeweils auf Basis der Intention, nach der eine Institution mit einer Qualitätseinschätzung beauftragt wurde: Zum Beispiel kann eine Institution den Auftrag erhalten, die Übereinstimmung von Vorgaben mit dem tatsächlich stattfindenden Prozess und dem Ergebnis zu überprüfen, während eine andere anhand quantitativer und qualitativer Methoden die Standards und Qualität an sich einschätzt. Auch können beide miteinander kombiniert werden, da nach beiden ein Interesse daran besteht, dass das Lehrpersonal ein Programm nach dem geltenden Standard gut vermittelt.

Diana Green versucht anhand eines Vergleichs verschiedener Qualitäts-Konzepte eine angemessene Definition für Qualität im Kontext Höherer Bildung anzubieten:

1. Zunächst beschreibt sie, was *Qualität als traditionelles Konzept* ausmache: Hierbei sei Qualität mit einer Auffassung verbunden, nach der ein Produkt oder eine Dienstleitung etwas Unverwechselbares und Spezielles sei, sowie

dem Besitzer Status verleihe. Darüber hinaus gälten extrem hohe Standards bei der Produktion, Auslieferung und Präsentation, die nur durch hohe Kosten sowie aufgrund knapper Ressourcen erreicht werden könnten. Aufgrund dessen werde die Bevölkerungsmehrheit ausgeschlossen und Exklusivität impliziert. Häufig werde hierfür der Rolls Royce als Beispiel herangezogen.

2. Eine weitere Definition beinhalte die Auffassung, dass Qualität mit der *Einhaltung einer Spezifikation (specification) oder eines Standards* verbunden sei: Ein Standard stünde für die Grundlage einer Messung oder eines Maßstabs, der eine erforderliche Eigenschaft eines Produktes beschreibe. Produkte und Dienstleitungen beinhalteten durch ihre Spezifikation eine Reihe von Standards, an deren Einhaltung Qualität gemessen werde.
3. Eine nach Green unter Analytikern und Politikern verbreitete Qualitätsdefinition bezüglich Höherer Bildung beschreibt den der *Zweckmäßigkeit (fitness for purpose)*: Hierbei habe Qualität keine Bedeutung außer in Bezug zu der Erfüllung des Zwecks eines Produkts oder einer Dienstleistung. Qualität werde an dem Ausmaß der Erfüllung dieses Zwecks beurteilt.

2.2 Was ist Qualität der Lehre?

Diese verschiedenen Perspektiven hinsichtlich des Verständnisses von Qualität, wurden von Diana Green (1994) hinsichtlich ihrer Übertragbarkeit auf den Bereich der Höheren Bildung diskutiert:

1. *Traditionelles Konzept:* Diese Perspektive auf Qualität sei für den allgemeinen Bereich der Höheren Bildung nicht hilfreich. Sie entspreche aber der Wahrnehmung vieler Menschen bezüglich der Universitäten Cambridge und Oxford.
2. *Einhaltung der Spezifikation oder Standards:* Laut Green passt diese Definition eher zu Qualität in der Höheren Bildung, da hierbei alle Institutionen die Möglichkeit hätten, Qualität zu erfüllen. Ein akademischer Standard werde in der Regel auf die Leistung von Studierenden bezogen. Ein breiterer Ansatz schließe das gesamte Spektrum von Aktivitäten im Bereich der Lehre und des Lernens und der Forschung im Hochschulbereich ein (unter anderem die Zulassungsverfahren, die Inhalte der Lehrveranstaltungen, die

Vermittlungsmethoden und die physischen Ressourcen). Jede Art von Institution könne sich unterschiedliche Standards setzen. Allerdings werde hierbei nichts über die Kriterien ausgesagt, die die Standards festlegen.

3. *Qualität als Zweckmäßigkeit:* Allgemein wurden im Kontext Höherer Bildung im Sinne dieser Zweckmäßigkeit die Vermittlung von Fertigkeiten (instruction in skills), die „Förderung der allgemeinen Verstandeskräfte“ (promotion of the general powers of the mind), die Förderung des Lernens (advancement of learning) und die Vermittlung einer gemeinschaftlichen Kultur und gemeinschaftlicher Normen einer Staatsbürgerschaft (transmission of a common culture and common standards of citizenship) genannt. Ebenso zähle hierzu auch die Deckung des Bedarfs der Wirtschaft, zum Beispiel in Form der angemessenen Ausbildung von Arbeitskräften oder ausreichend hoher Absolventenzahlen. Eine weitere Version von Zweckmäßigkeit beziehe sich auf eine Qualität, die die Erfüllung von Kundenwünschen beinhalte. Hierbei wäre zu klären, wer in der Höheren Bildung der Kunde sei: die Studierenden, die Angestellten oder die Regierung? Zusätzlich seien Bedürfnisse schwer feststellbar, und es bestehe die Frage, ob Studierende ihre Bedürfnisse identifizieren könnten.

   Eine hochqualitative Bildungseinrichtung zeichne sich dadurch aus, dass sie ihren Zweck klar benenne und die selbst gesteckten Ziele effektiv und effizient erreiche.

   Das Problem dieser Definition bestehe darin, dass nicht klar sei, welchem Zweck Höhere Bildung konkret diene. Die verschiedenen genannten Zwecke könnten auch miteinander kollidieren.

Als Lösung des Problems der Definition von Qualität in der Höheren Bildung schlägt Green eine pragmatische Vorgehensweise vor (S. 17): Jeder Akteur solle seine verwendeten Qualitätskriterien klar benennen, so dass diese bei Vergleichen berücksichtigt werden könnten. Denn es gebe kein einheitliches Qualitätskonzept, da jede Gruppe oder Institution andere Prioritäten habe. Beispielsweise liege für Studierende und Dozenten der Aufmerksamkeitsfokus auf dem Bildungsprozess, während er für Arbeitgeber auf dem „Output“ liege.

Ähnlich wie Diana Greens Ausführungen, hält der Wissenschaftsrat (2008) in den „Empfehlungen zur Qualitätsverbesserung von Lehre und Studium“ fest, dass das

Verständnis von Qualität der Lehre an Kontexte gebunden sei. Dieses Qualitätsverständnis werde daran bemessen, welche Ziele und Wirkungen mit Lehre und Studium jeweils verbunden seien und in welchem Ausmaß diese Vorgaben erreicht und umgesetzt würden. Damit gebe es unterschiedliche Auffassungen davon, was der Zweck von Hochschulen und - damit verbunden - was das Verständnis von Qualität der Hochschullehre sei:

- Für Studierende und Arbeitgeber sei Qualität die angemessene Vorbereitung auf eine berufliche Tätigkeit.
- Hochschullehrer verstünden unter Lehrqualität die Vermittlung des wissenschaftlich abgesicherten Erkenntnisfortschritts.
- Staat und Geldgeber sowie die Öffentlichkeit wiederum betrachteten vor allem Verlässlichkeit und Aussagekraft von Studien- und Abschlussniveau als maßgebliche Qualitätskriterien.
- Die Erwartungshaltung der Politik und Öffentlichkeit beinhalte, dass Hochschulabsolventen in der Lage seien, die kulturelle, soziale, technologische und wirtschaftliche Weiterentwicklung der Gesellschaft verantwortungsvoll voranzubringen.

Diese verschiedenen Auffassungen bildeten ein sich nicht gegenseitig ausschließendes, komplexes, multidimensionales und multifunktionales, die vielfältigen Aspekte der Hochschulbildung berücksichtigendes Qualitätsverständnis.

2.3 Schlussfolgerungen

Wie beschrieben stellt der Wissenschaftsrat (2008) fest, dass das Verständnis von Qualität der Lehre an Kontexte gebunden sei. Qualität bemesse sich daran, welche Ziele und Wirkungen mit Lehre und Studium jeweils verbunden seien und in welchem Ausmaß Vorgaben erreicht und umgesetzt würden. Was bedeutet dies übertragen auf die Ergebnisse studentischer Lehrevaluationsinventare?

In dieser Arbeit wird angenommen, dass sich das Qualitätsverständnis von Studierenden hinsichtlich der Lehre maßgeblich auf die Deckung des eigenen Bedarfs hinsichtlich verschiedener Aspekte des Studiums bezieht. Entsprechend nennen Westermann, Spies, Heise & Wollburg-Claar (1998) als zu messende Konstrukte in der Evaluation der Lehre die *Zufriedenheit oder Unzufriedenheit mit einer bestimmten*

Lehrveranstaltung und die *Zufriedenheit oder Unzufriedenheit mit veranstaltungsübergreifenden Studienbedingungen.* Letztere betrifft beispielsweise Aspekte des Bibliotheksangebots oder der Studienberatung.

In dieser Arbeit werden darauf aufbauend studentische Lehrevaluationsergebnisse als Maß der studentischen Zufriedenheit mit einer bestimmten Lehrveranstaltung behandelt. Konkreter bedeutet dies, dass anhand eines studentischen Lehrevaluationsergebnisses erfasst werden soll, in welchem Ausmaß Lehrqualität in dem Sinne gegeben ist, in dem Studierende mit der Durchführung einer bestimmten Lehrveranstaltung und der Vermittlung von Lehrinhalten zufrieden sind. Die Zufriedenheit mit veranstaltungsübergreifenden Bedingungen soll nicht Gegenstand dieser Arbeit sein.

Nach Green (1994) solle jeder Akteur seine Qualitätskriterien klar benennen. Somit müsse für die Gestaltung eines studentischen Lehrevaluationsinventars auch geklärt sein, welche Kriterien hinsichtlich Lehrqualität im Sinne der Durchführung einer Lehrveranstaltung und Vermittlung von Lehrinhalten erfüllt sein sollten, um von einem gedeckten Bedarf der Studierenden ausgehen zu können.

3. Facetten von Lehrqualität

Der Beantwortung der Frage, welche Kriterien von Lehrqualität im Sinne der Deckung des Bedarfs von Studierenden hinsichtlich der Durchführung einer Lehrveranstaltung und Vermittlung von Lehrinhalten relevant sind, kann sich anhand verschiedener Quellen angenähert werden: Was Lehrqualität aus wissenschaftlich-didaktischer Sicht ausmacht, kann aufgrund theoretischer Überlegungen und durch empirische Studien verschiedenster Art untersucht und begründet werden. Zu letzterem gehören maßgeblich Befragungen von Lehrenden und Studierenden, Beobachtungen und Beschreibungen realer Veranstaltungen oder die Auswertung von Lehrevaluationsinventaren. Um einen Überblick zu bieten, werden im Folgenden zunächst aus dieser Vielzahl von Ansätzen allgemeine Empfehlungen, Modelle der Lehr- und Unterrichtsforschung, Studierendenbefragungen sowie allgemeine Schlussfolgerungen aus der Sichtung verschiedener Studienergebnisse vorgestellt. Diese können als theoretischer Hintergrund für die Konstruktion eines Lehrevaluationsinventars dienen. Im späteren Verlauf dieser Arbeit werden diese noch durch weitere Ansätze ergänzt (siehe Kapitel 8).

3.1 Empfehlungen universitärer Institutionen

Verschiedene Institutionen haben allgemeine Richtlinien aufgestellt, an denen sich zur Förderung von Lehrqualität orientiert werden kann. Im Folgenden wird exemplarisch eine fächerübergreifende Ausarbeitung der Johannes Gutenberg-Universität Mainz sowie ein auf die medizinische Ausbildung zugerichtetes Programm beschrieben.

3.1.1 Aspekte guter Lehre

Die Johannes Gutenberg-Universität in Mainz versteht Qualität der Lehre „als Maß der Übereinstimmung von Lehrzielen und Lehrpraxis unter der Maßgabe, dass ein Abgleich zwischen Teilzielen bzw. zwischen über- und untergeordneten Zielen erfolgt“. Sie hat auf Basis der Empfehlungen des Wissenschaftsrats (Wissenschaftsrat, 2008) „Aspekte guter Lehre“ zusammengestellt: (Zentrum für Qualitätssicherung und -entwicklung, 2011, S. 3)

1. *Mehrdimensional:* Lehre sei mehrdimensional und müsse gegenläufigen Ansprüchen gerecht werden. Mit letzterem ist beispielsweise die Vermittlung allgemeiner Kenntnisse als auch spezifischer Fachkenntnisse gemeint oder die

Einbindung von Studierenden in den Forschungsprozess bei gleichzeitiger Beachtung derer, die keine wissenschaftliche Ausbildung anstreben.

2. *Fachverständnis:* Ein Fachverständnis solle die Frage beantworten, was ein Fach unter gegeben Rahmenbedingungen leisten könne. Darunter fielen unter anderem Schwerpunktsetzungen, Kooperationen mit und Grenzen gegenüber anderen Fächern sowie Übereinkünfte in Hinblick auf wissenschaftliche Fachstandards. Die Definition eines Fachverständnisses sei Grundlage der Festlegung von Lernzielen.
3. *Lernziele:* Lernziele würden auf Basis eines jeweiligen Fachverständnisses abgeleitet und an Studierende kommuniziert.
4. *Anschlussfähigkeit:* Lehrveranstaltungen und Studienabschnitte sollten aufeinander aufbauen und auf die erwarteten Anforderungen nach dem Studium abgestimmt sein.
5. *Gute Betreuung von Studierenden:* Diese sei durch angemessene und frühzeitige Leistungsrückmeldung sowie der Förderung des Potenzials von Studierenden charakterisiert.
6. *Forschungsleistung:* Ein hohes Forschungsniveau führe zu hohem Lehrniveau.
7. *Weiterbildung:* Fertigkeiten für gute Lehre sollen erworben und weiterentwickelt werden.
8. *Verständnis von guter Lehre:* Solch ein Verständnis differiere zwischen verschiedenen Fächern, Fächergruppen und Studiengängen. Somit könnten Kriterien von Lehrerfolg eine sehr unterschiedliche Gewichtung erfahren.
9. *Fachspezifische Indikatoren:* Die Messbarkeit von Effekten guter Lehre hinsichtlich des Fachverständnisses und der damit verbundenen Zielsetzungen werde als gegeben angenommen und die Einschätzung der Studierenden sei einzubeziehen. Kriterien guter Lehre seien ein erfolgreicher Studienabschluss und eine fachnahe Berufseinmündung. Fachstudiendauer, Daten zum Studienverlauf und Prüfungsergebnisse hätten eine Relevanz bezüglich Lehr- und Lerneffekten, wären aber unter intervenierenden Variablen wie dem Leistungsvermögen der Studierenden und der Situation auf dem Arbeitsmarkt zu interpretieren. Fachspezifische Erfolgskriterien, die im Zeitvergleich Rückschlüsse auf die Lehrleistungen des Fachs zulassen, sollten entwickelt werden.

3.1.2 Das Stanford Faculty-Programm

Als Beispiel für fachspezifische Aspekte guter Lehre kann im medizinischen Kontext das *Stanford Faculty Development Program (SFDP)* (Skeff, Stratos, Berman & Bergen, 1992) angeführt werden: Das SFDP ist ein Programm zur Verbesserung klinischer Lehre, dessen konkrete Inhalte auf der Forschung zu klinischer Lehre im Krankenhaus als auch im ambulanten Bereich basieren. Es wird Dozenten medizinischer Fakultäten in Seminarform angeboten und dabei werden diese auch darin geschult, das erworbene Wissen an ihre Kollegen weiterzugeben. Inhaltlich wird der Fokus auf die Vermittlung klinischer Themen gesetzt (beispielweise Themen der ambulanten Versorgung), die Art der Vermittlung entspricht aber allgemeinen Grundsätzen von Lehre und lässt sich in sieben Kategorien zusammenfassen (Skeff, 1988):

1. *Etablierung eines positiven Lernklimas:* Die Lernenden sollen sich wohl und angeregt fühlen.
2. *Leitung einer Lerneinheit:* Der Lehrende soll fähig sein, effektiv die Lehreinheit zu managen, zu fokussieren und zeitlich anzupassen.
3. *Zielkommunikation:* Der Lehrende soll die beabsichtigten Ergebnisse hinsichtlich Fähigkeiten, Einstellungen und Wissen klar aufstellen, aussprechen und die Erwartungen besprechen.
4. *Förderung von Verstehen und Behalten:* Es sollen Lehrmethoden eingesetzt werden, die das anfängliche Begreifen und das Erinnern des entsprechenden Lerninhalts fördern.
5. *Evaluation:* Durch Methoden des Lehrenden soll eingeschätzt werden, ob die erwünschten Lernziele durch die Lernenden erreicht wurden. Damit kann der weitere Verlauf geplant aber auch die abschließende Kompetenz eingeschätzt werden.
6. *Feedback:* Der Lehrende soll die Lernenden darüber informieren, wie sie ihre Leistung verbessern könnten.
7. *Förderung selbstbestimmten Lernens:* Der Lehrende soll die Fähigkeiten der Lernenden darin fördern, ihren eigenen Lernbedarf ohne oder mit der Hilfe anderer zu identifizieren und entsprechend zu handeln.

Das Programm wurde von seinen Teilnehmern wie auch von Fakultäten als sehr nützlich beurteilt. Mitarbeiter und Studierende berichteten von einer verbesserten

Lehrdurchführung (Skeff et al., 1992). Weithin wurden auf das Stanford Faculty Development Program bezogene studentische Lehrevaluationsbögen entwickelt (Litzelman, Stratos, Marriott & Skeff, 1998) und auch in Deutschland etabliert (Iblher et al., 2011).

3.2 Modelle und Theorien zur Lehr- und Unterrichtsqualität

Es gibt eine Vielzahl theoretischer Ansätze und Modelle, die thematisierten beziehungsweise untersucht haben, was Lehrqualität ausmacht. Im Folgenden werden zwei dieser Ansätze vorgestellt, die in der Lehr- und Unterrichtsforschung verbreitet sind: Das *Constructive Alignment* und das *Angebots-Nutzungsmodell* von Helmke.

3.2.1 Constructive Alignment

Constructive Alignment ist ein didaktisches Konzept für Lehr- und Lernsituationen, das auf soziokulturelle und linguistische Schulen des Konstruktivismus zurückgreift. In diesen wird sich auf Kontexte und Wege bezogen, anhand derer das Bewusstsein Wissen konstruiert. (Biggs, 1996)

Grundlegend für alle konstruktivistischen Theorien sei, dass Lernende durch eine aktive Selektion, kumulative Konstruktion und ihrem eigenen Wissen anhand individueller und sozialer Aktivität „Bedeutung“ konstruieren. Der Lernende bringe eine Akkumulation von Voraussetzungen, Motiven, Intentionen und Vorwissen mit, die jede Lehr-Lern-Situation beinhalte und die Entwicklung der Qualität des Lernens determiniere.

Bezüglich der Anwendung werde nicht eine konstruktivistische Methode vermittelt, sondern eine Lehr-Einstellung, die ein fokussiertes Bewusstsein hinsichtlich des Lernenden und seiner Welt impliziere (Martin & Booth, 1996; zitiert nach Biggs, 1996, S. 349). Lehren forme laut Biggs (1993) ein komplexes System aus den Lehrenden, den Lernenden, dem Lehrkontext, den Lernaktivitäten der Lernenden und dem Ergebnis (Outcome). Dieses System sei innerhalb eines größeren institutionellen Systems eingebettet. Innerhalb dieses Systems sollten alle Elemente aufeinander ausgerichtet sein:

1. Lehrende sollten sich im Klaren darüber sein, was sie den Lernenden beibringen möchten („intended learning outcomes“), und wie sich das Lernen

in Verstehensleistungen manifestieren solle (zum Beispiel das Wiedererkennen des Gelernten in einem neuen Kontext statt reinem Erinnern).

2. Die Perfomance-Ziele werden in einem hierarchischen Bewertungssystem von „höchst akzeptabel“ zu „kaum befriedigend“ eingestuft.
3. Lernende sollen in Situationen gebracht werden, die als geeignet angesehen werden, das Gelernte hervorzubringen.
4. Die Studierenden sollten dann Evidenzen dafür Erbringen, dass ihr Lernen mit den festgelegten Zielen übereinstimmt.

Zusammengefasst beinhaltet *Constructive Alignment* drei Kernelemente, die aufeinander ausgerichtet und voneinander abhängig sind: Eine Lehrveranstaltung sollte so gestaltet sein, dass die Lernenden durch die Lehr-Lernaktivität die angestrebten Ziele auch erreichen können und eine Prüfung das Erreichen genau dieser Ziele testet (siehe Abbildung 1).

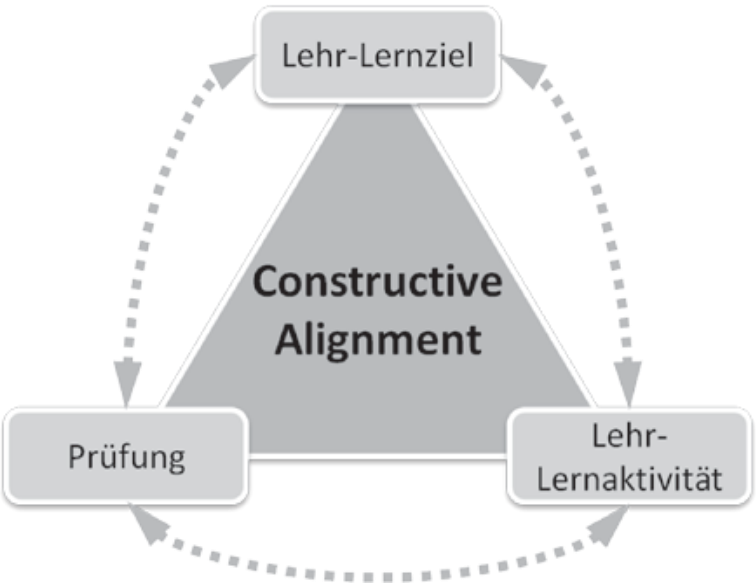

Abbildung 1: Die Kernelemente des Constructive Alignments und ihre Beziehung zueinander (Abbildung aus Baumert & May, 2013, S. 23)

3.2.2 Angebots-Nutzungs-Modell von Helmke

Helmke (2006) beschreibt das von ihm entwickelte *Angebot-Nutzungs-Modell* als Ausdruck des gegenwärtigen Wissens über Bedingungen, Vernetztheit und Konsequenzen von Unterricht (siehe Abbildung 2). Es bezieht sich auf schulischen Unterricht, kann aber auch relevante Aspekte für die universitäre Lehre aufzeigen.

Allgemein sagt dieses Modell aus, dass die Wirksamkeit eines Unterrichtsangebots von der Nutzung abhänge und dass je nach Bildungsziel verschiedene Lehr-Lern-Arrangements notwendig sein könnten. Für guten Unterricht seien drei Perspektiven sinnvoll: Die Lehrkompetenz der Lehrperson, die Qualität der Unterrichtsprozesse und die entsprechenden Effekte. Diese Konstellation werde von Rahmenbedingungen mitgeprägt (beispielsweise, ob eine Schule in einem sozialen Brennpunkt liege oder Schüler mit geringem Vorkenntnisniveau in der Klasse seien oder welches Fach unterrichtet werde).

Eine weitere grundlegende Annahme sei, dass Wirkungsaussagen auf Wahrscheinlichkeiten beruhten, da das Gesamtprofil im Kontrast zu einzelnen Aspekten höhere Wirkungsgrade habe.

Helmke fasst zehn Merkmale eines fächerübergreifenden guten Unterrichts zusammen, wobei Punkt 10 als Schlüsselmerkmal aufgefasst wird:

1. *Effiziente Klassenführung und Zeitnutzung:* Die Etablierung und Einhaltung von Regeln.
2. *Lernförderliches Unterrichtsklima:* Freundlichkeit, Humor, Respekt und so wenig Lernsituationen mit Leistungsbewertung wie möglich.
3. *Vielfältige Motivierung:* Die Thematisierung unterschiedlicher lernrelevanter Motive und die Anregung des Neugier- und Leistungsmotivs.
4. *Strukturiertheit und Klarheit:* Angemessene Sprache, strukturierende Hinweise wie Vorschauen, fachlich-inhaltliche Korrektheit und sprachliche Prägnanz.
5. *Wirkungs- und Kompetenzorientierung:* Ein Fokus auf den Erwerb fachlicher, überfachlicher und nichtfachlicher Kompetenzen sowie auf nachweisliche und nachhaltige Wirkungen.
6. *Schülerorientierung, Unterstützung:* Lehrkräfte sollten fachliche und persönliche Ansprechpartner sein. Lernende sollten angemessen mitbestimmen und Feedback abgeben können.
7. *Förderung aktiven, selbstständigen Lernens:* Förderung von selbstständigem, eigenverantwortlichem Lernen mit Sprech- und Lerngelegenheiten für alle Schüler.
8. *Angemessene Variation von Methoden und Sozialformen:* Schüler-, fach- und lernzielangemessene Variationen.

9. *Konsolidierung, Sicherung, intelligentes Üben:* Vielfalt an Aufgaben und Bereitstellung unterschiedlicher Transfermöglichkeiten.
10. *Passung:* Zum Beispiel die Anpassung der Schwierigkeit an die jeweilige Lernsituation und die Lernvoraussetzungen der Schülergruppen beziehungsweise der Umgang mit Heterogenität. Laut Helmke das Kernmerkmal, da es den Umgang mit Heterogenität beinhalte und für alle Lehr-Lernsituationen gültig sei.

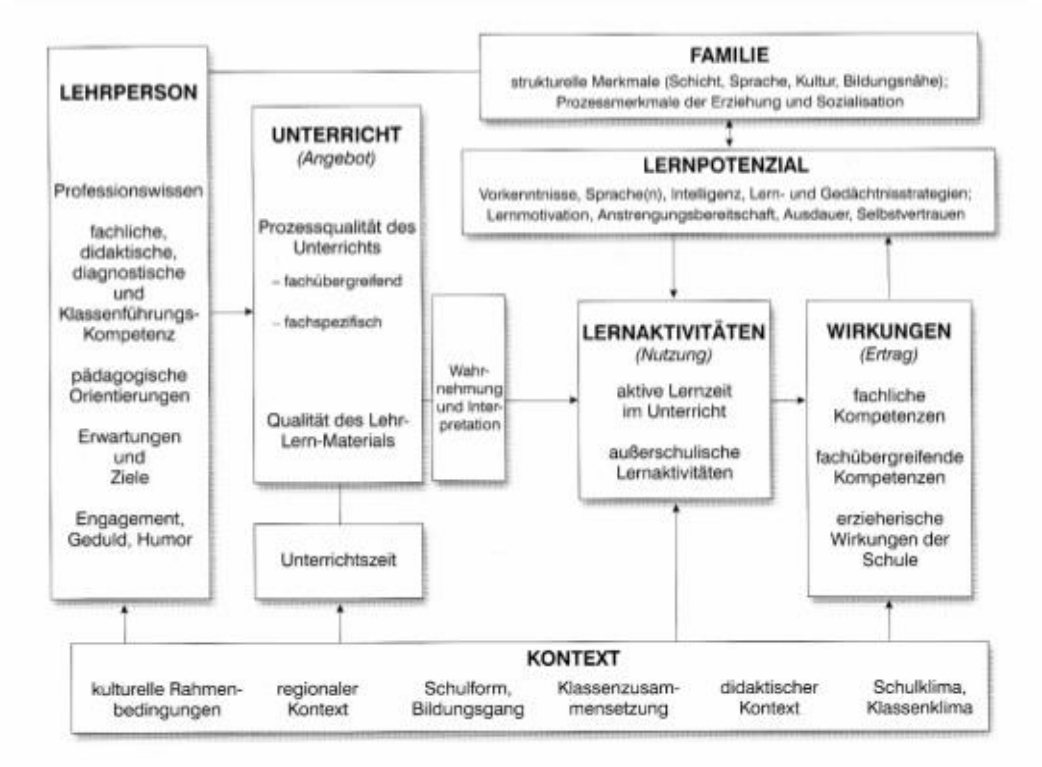

Abbildung 2: Das Angebots-Nutzungs-Modell nach Helmke (Abbildung aus Helmke, 2009, S. 73)

3.3 Studierendenbefragungen

3.3.1 „The superior college teacher from the students‘ view”

Feldman (1976) hat sich mit der Frage beschäftigt, welche Einstellungen und Verhaltensweisen mit herausragenden (superior) College-Dozenten assoziiert sind. Grundlage von Feldmans Arbeit ist eine Zusammenstellung der damals (1976) vorhandenen Forschung an nordamerikanischen Studierenden.

Anhand von drei Vorgehensweisen wurde die Fragestellung beantwortet:

1. Studierende sollten den für sie idealen Dozenten beschreiben.
2. Studierende sollten aufführen, was sie glauben welche Charakteristika besonders wichtig für gute Lehre seien.
3. Studierende sollten die besten Dozenten beschreiben, die sie bislang gehabt hatten.

Feldman schloss 49 Studien in seine Zusammenstellung ein und es zeigten sich fünf Charakteristika, die konsistent mit herausragenden Dozenten assoziiert wurden:

1. *Anregung von Interesse:* Beinhaltet zum Beispiel, dass der Dozent intellektuelle Neugier wecken konnte, und es somit leicht war, aufmerksam zu bleiben.
2. *Klarheit und Verständlichkeit:* Zum Beispiel waren die Erläuterungen des Dozenten verständlich, oder er nutzte gute Beispiele und Illustrationen für schwierige Aspekte.
3. *Das Wissen des Dozenten über den Unterrichtsgegenstand:* Der Dozent hatte ein gutes Wissen über den Lehrbuchinhalt beziehungsweise ein vollständiges Wissen über den Unterrichtsgegenstand.
4. *Die Vorbereitung des Dozenten und die Organisation des Kurses:* Zum Beispiel, dass der Dozent für jede Vorlesung gut vorbereitet war.
5. *Enthusiasmus des Dozenten für das Thema und für die Lehre:* Der Dozent hatte beispielsweise Spaß an der Lehre.

In der freien Beschreibung der Studierenden hinsichtlich der Frage, welche Charakteristika gute Lehre ausmache und bei der Beschreibung des idealen Dozenten, wurden drei weitere Aspekte identifiziert:

1. *Freundlichkeit des Dozenten, Sorge und Respekt für Studierende:* Zum Beispiel, dass der Dozent Studierende ernst nahm und zu allen Studierenden freundlich war.
2. *Verfügbarkeit und Hilfsbereitschaft des Dozenten:* Der Dozent war bereit, Studierenden bei Schwierigkeiten zu helfen beziehungsweise, dass der Dozent für eine Beratung erreichbar war.
3. *Ermunterung zu Fragen und Diskussionen, Offenheit für die Meinung von anderen:* Der Dozent regte die Diskussionen in dem Kurs an beziehungsweise forderte zu Kritik an seinen eigenen Ideen auf.

Diese scheinen aber für die Gesamtbeurteilung von Dozenten nicht wichtig zu sein, da sie bei einer vorgegebenen Liste von Eigenschaften als weniger wichtig angegeben wurden. Feldman zieht die Möglichkeit in Betracht, dass diese Eigenschaften als allgemeine Annahme eines Verhaltensrepertoires angesehen werden, aber je spezifischer und strukturierter die Situation würde, desto wichtiger würden andere Annahmen.

3.3.2 Charakteristika effektiver College-Dozenten

Onwuegbuzie et al. (2007) befragten Studierende an einem College, welche Charakteristika sie bei effektiven (effective) College-Dozenten wahrnehmen. Dafür sollten sie zwischen drei und sechs Charakteristika angeben und für jede eine Definition oder Beschreibung hinzufügen.

Es nahmen 912 Studierende daran teil, und als Ergebnis stellten sich neun Themen heraus:

1. *Eingehend (responsive):* zum Beispiel, dass der Dozent den Studierenden Rückmeldung über die Leistung gibt
2. *Begeistert (enthusiast):* zum Beispiel, dass bei dem Dozenten Leidenschaft für das unterrichtete Fach vorhanden ist
3. *Auf die Studierenden zentriert (student centered):* zum Beispiel auf Probleme von Studierenden einzugehen und starke zwischenmenschliche Kompetenzen zu besitzen
4. *Professionell:* zeigt Verhaltensweisen und Veranlagungen, die vorbildlich für die Fachrichtung eines Dozenten sind (displays behaviors and dispositions deemed exemplary for the instructor's discipline); zum Beispiel wurden Ziele gesetzt, die zu erreichen sind
5. *Experte:* der Dozent besitzt ein Wissen über den Kursinhalt und darüber hinaus
6. *Verbindend (connector):* der Dozent ist für Studierende erreichbar und kann dadurch zusätzliche Hilfe anbieten
7. *Übermittler/Vermittler (Transmitter):* kann das Interesse der Kursteilnehmer aufrechterhalten, hat gute sprachliche Qualitäten
8. *Moralisch (ethical):* behandelt alle Studierenden gleich

9. *Leiter (Director):* bietet eine sichere und geordnete Lernumgebung durch effiziente Zeitstruktur und optimierte Ressourcen

3.4 Weitere Studienergebnisse

Rindermann (2009, S. 55-56) kommt nach der Sichtung vieler verschiedener Studien zusammenfassend zu dem Schluss, dass insbesondere folgende drei Aspekte relevant für gute Lehre seien:

1. Gute Strukturierung
2. Didaktische Methodenvielfalt und –sicherheit
3. Soziale Kompetenz und Persönlichkeitseigenschaften wie Freundlichkeit, Offenheit und Engagement.

3.5 Schlussfolgerungen

In diesem Kapitel wurden verschiedene Empfehlungen, theoretische Ansätze und Studienergebnisse vorgestellt. Ihnen allen ist gemein, dass sie durch die jeweils genannten Kriterien Lehrqualität beschreiben beziehungsweise definieren wollen. Dabei wurden mehrere Kriterien über verschiedene Ansätze und Studienergebnisse hinweg mehrmals genannt (zum Beispiel „Klarheit“ und „Verständlichkeit“).

Wie in Kapitel 2 geschlussfolgert, kann sich Qualität der Lehre auf verschiedene Kontexte beziehen. In den in diesem Abschnitt genannten Veröffentlichungen wurden verschiedene Begriffe wie „gute oder effektive Lehre“ beziehungsweise „effektive oder herausragende Dozenten“ verwendet. In dieser Arbeit soll Lehrqualität in dem Sinne abgebildet werden, nach der die Zufriedenheit Studierender anhand relevanter Aspekte hinsichtlich der Durchführung einer Lehrveranstaltung und der Vermittlung von Lehrinhalten gegeben ist. Somit kann es sein, dass nicht alle der in diesem Kapitel genannten Kriterien für diese Sicht auf Lehrqualität relevant sind und dementsprechend nicht anhand eines Lehrevaluationsinventars abgefragt werden sollen.

Unabhängig davon, zu welchem Zweck und anhand welcher Kriterien Lehrqualität erfasst werden soll, wird sie idealerweise in einem Evaluationsprozess systematisch überprüft und anhand von Messinstrumenten erfasst. Im folgenden Kapitel wird dies unter dem Begriff der *Evaluation* zusammengefasst und an Beispielen erläutert.

4. Wie wird Lehrqualität gemessen?

Um zu überprüfen, in welchem Ausmaß Qualität der Lehre jeglicher Art erfüllt ist, muss diese gemessen werden. Wie beschrieben ist dabei zunächst zu beachten, zu welchem Zweck Lehrqualität gemessen werden soll. Auf dieser Basis kann entschieden werden, anhand welcher Daten eine Erhebung mit anschließender Beurteilung durchgeführt werden sollte.

Allgemein kann solch eine systematische Qualitätsbeurteilung als *Evaluation* bezeichnet werden. Im Folgenden soll der Begriff der Evaluation definiert und der Ablauf eines Evaluationsprozesses skizziert werden. Im Anschluss werden beide Konzepte auf die Erfassung von Lehrqualität in Form von Lehrevaluationen übertragen und bereits konstruierte Inventare vorgestellt.

4.1 Lehrevaluation

4.1.1 Was versteht man unter Evaluation?

Evaluation an sich ist ein in verschiedenen Themenspektren eingesetztes Verfahren. Moosbrugger und Schweizer (2002, S. 20) definieren Evaluation allgemein als Überprüfung von Maßnahmen im Sinne einer Bewertung: Sie beurteile oder behaupte den Nachweis der Wirksamkeit einer Maßnahme, aber ohne diese Wirksamkeit zu erklären. Neben diesem gebe es auch den Begriff der *Evaluationsforschung*: Dieser wird als Optimierung der Überprüfung von Maßnahmen charakterisiert. In einer weiteren Definition wurde *Evaluationsforschung* von Rossi und Freeman (1993, zitiert nach Bortz & Döring, 2006, S. 96) als „die systematische Anwendung empirischer Forschungsmethoden zur Bewertung des Konzeptes, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme“ beschrieben. Evaluationen werden hinsichtlich Erkenntnis, Optimierung, Kontrolle, Entscheidungen und Legitimation von Maßnahmen oder Interventionen eingesetzt (Bortz & Döring, 2006, S. 97).

Evaluationen können *summativer* oder *formativer* Art sein (Bortz & Döring, 2006, S. 109): Erstere wird angewandt, wenn eine vorgegebene Maßnahme abgeschlossen und zu beurteilen ist. Bei letzterer werden regelmäßig Zwischenergebnisse erstellt und diese zur Modifikation oder Verbesserung dieser Maßnahmen verwendet.

4.1.2 Je nach Zweck und Verwendung unterschiedliche Evaluationsformen

Eine summative Evaluation hinsichtlich der Lehrqualität kann untersuchen, ob eine Lehrveranstaltung von Studierenden angenommen wird und überhaupt fortgeführt werden soll. Eine formative Evaluation kann dazu eingesetzt werden, auf den jeweils aktuellen Bedarf von Studierenden einzugehen oder dem Dozenten und den Organisatoren kontinuierlich Rückmeldung geben zu können.

Um Lehrqualität messen zu können, können je nach Zweck verschiedene Formen der Lehrevaluation eingesetzt werden:

- Um den Lehr- oder Lernerfolg zu messen, dienen beispielsweise quantitative Indikatoren in Form der Abschlussnote.
- Um zu erfassen, inwiefern eine Universität die Nachfrage nach Menschen mit bestimmten fachlichen Abschlüssen bedient, können quantitative Indikatoren in Form von Abschlusszahlen abgerufen werden.
- Um zu beurteilen, ob ein Dozent den Studierenden einen angemessenen Überblick über den Inhalt seines Fachgebiets vermittelt, können systematische Beurteilungen durch Experten erfolgen.
- Zur Beurteilung, inwiefern ein Dozent eine Lehrveranstaltung angemessen durchführt und Lehrinhalte vermittelt, können Selbsteinschätzungen, systematische Beobachtungen von Experten sowie studentische Lehrevaluationen eingesetzt werden.

4.1.3 Studentische Lehrevaluationen

Rindermann (2009, S. 26) empfiehlt, dass Lehrende und Wissenschaftler die Lehrinhalte bewerten sollten und Hochschuldidaktiker sowie Studierende die Vermittlung dieser Inhalte. Studierende seien aber gegenüber den Didaktikern im Vorteil, da sie Veranstaltungen häufiger besuchten, dadurch Vergleiche zwischen Lehrenden ziehen könnten und letztendlich auch die Adressaten von Lehre seien.

Studierende könnten Lehrveranstaltungen in Form von Befragungen oder Fragebögen evaluieren. Fragebögen seien allerdings aufgrund ihrer ökonomischen Verwendung und der Überprüfbarkeit ihrer Güte und Normierung im Vorteil. Nachtteile lägen jedoch in der Vorgegebenheit von Items, Dimensionen und Antwortskalen sowie der Verfälschbarkeit durch Antwortstile, wie der *Ja-Sage-Tendenz*. Rindermann. (2009, S. 59)

Im Laufe der Zeit wurden in Deutschland und international eine große Menge studentischer Lehrevaluationsinventare konstruiert und eingesetzt: Zu den ersten in Deutschland verwendeten Inventaren für die Rückmeldung von Studierenden zählt laut Souvignier und Gold (2002, S. 226) beispielsweise das 1971 von Müller-Wolf und Fittkau (1971) publizierte. Im weiteren zeitlichen Verlauf vervielfältigte sich die Anzahl publizierter und nicht publizierter Inventare, die auch für unterschiedliche Zwecke konstruiert wurden und somit zumindest teilweise unterschiedliche qualitätsrelevante Aspekte abfragen: Beispielsweise ist das *Heidelberger Inventar zu Lehrveranstaltungsevaluation* (HILVE-2) zur Evaluation von Lehrveranstaltungen jeglicher Art vorgesehen (Electric Paper - Gesellschaft für Softwarelösungen, 2004), andere wiederum spezifisch für Seminare, Vorlesungen oder Praktika (Staufenbiel, 2000). Weitere sind auf das Verhalten des Dozenten spezialisiert (Koch, 2004) oder auf spezifische Lehrinhalte wie Unterstützungsprogramme für Promovierende (Paulitsch et al., 2016).

Es gibt eine Diskussion darüber, wie Lehrevaluationsinventare aufgebaut sein sollten, beziehungsweise wie die Urteile der Studierenden besser abgefragt werden könnten. Ist eine mehrdimensionale Struktur vorteilhafter, da somit verschiedene Kriterien erfasst werden, oder sind globale Urteile wegen einer potenziell geringeren Anfälligkeit für Verzerrungen angemessen?

Marsh (1987, 2007) schlägt den Einsatz von *Factor Scores* vor, die auf Basis von Faktorenanalysen abgeleitet wurden (siehe Kapitel 5.2.3 über Faktorenanalysen). Dem widerspricht Abrami (1989) und schlägt für summative Zwecke - insbesondere bei Beförderungen und Einstellungs-Entscheidungen - mehrere globale Urteile vor (zum Beispiel, die allgemeine Einschätzung der Fähigkeiten des Dozenten) oder vorsichtig gewichtete, gemittelte Rating-Faktoren (weighted average of rating factors). Als Begründung führt Abrami folgende Aspekte auf:

1. Das Fehlen guter Theorien darüber, was gute Lehre ausmacht und inkonsistente Ergebnisse von Faktorenanalysen.
2. Die Inhaltsvalidität spezifischer Items hinsichtlich verschiedener Kurse, Dozenten, Studierenden und Settings wird bezweifelt: zum Beispiel könne bei kleinen und großen Kursen die Frage, ob der Dozent zum einzelnen Studierenden freundlich war, unterschiedlich relevant sein.

3. Laut einem Review von Cohen (1981, zitiert nach Abrami, 1989, S. 223) hätten in *Multisection Validity Studies*[2] globale Beurteilungen des Kurses oder des Dozenten höhere Korrelationen mit *studentischem Lernen* (student learning) als viele der einzelnen Dimensionen.
4. Man wisse weniger über die Generalisierbarkeit von Beurteilungen einzelner Dimensionen als bei globalen Ratings in Anbetracht einer Vielzahl von Bedingungen (verschiedene Kurse, Dozenten, Studierenden und Situationen).
5. Man könne von Nicht-Experten oder Angehörigen der Universitätsleitung (Administrators) nicht erwarten, die einzelnen Dimensionen zu gewichten, um auf deren Basis eine Entscheidung hinsichtlich der Lehrqualität eines Dozenten zu fällen. Aus persönlicher Erfahrung des Autors (Abrami) gewichteten Leitungsangehörige die einzelnen Dimensionen gleich.
6. In einem späteren Artikel (Abrami, d'Apollonia & Rosenfield, 1997) wird noch ergänzt, dass bei gut konstruierten multidimensionalen Inventaren die globalen Items meist stark auf die ersten Faktoren laden würden, und die Menge unterschiedlicher Inventare ein Zeichen dafür sei, dass es keinen klaren Konsens gebe, aus welchen Dimensionen Lehrevaluation bestehe.

Für formative Zwecke hält Abrami (1989) allerdings globale Ratings weitgehend für ungeeignet. Weiterhin stimmt er mit Marsh (1987) überein, dass dimensionale Ratings validiert werden müssten, wenn sie nur zur Verbesserung der Lehre eingesetzt würden.

Marsh (2007) kritisiert den Vorzug globaler Ratings gegenüber den Faktor-Scores in folgender Hinsicht:

1. Er bevorzuge nicht nur Factor-Scores, sondern ein Profil an Werten, darunter auch globale Urteile.

[2] In *Multisection Validity Studies* werden Studierende zufällig verschiedenen Abschnitten (Sections) zugewiesen, um anfängliche Unterschiede zu verringern. Jeder Abschnitt hat einen anderen Dozenten, ansonsten sind Lehrinhalte, Lehrmaterialen und Abschlussprüfung gleich. Vor der Bekanntgabe der Abschlussnote evaluieren die Studierenden die Dozenten.

2. Auf das Argument, dass es keinen Konsens darüber gebe, welche Dimensionen erhoben werden sollten, entgegnet Marsh, dass die in empirischer Forschung identifizierten Dimensionen der Lehrevaluation schon von Feldmann (1976) umfassend dargestellt wurden (siehe Abschnitt 3.3.1) und in jedem Instrument vorkämen.
3. Studentisches Lernen korreliere in *Multisection Validity Studies* mit spezifischen Dimensionen systematisch höher als mit globalen Ratings: Ergebnisse zeigten, dass in den Abschnitten (Sections), in denen die Lehre als am effektivsten evaluiert wurde, auch die Abschlussprüfungen besser ausfielen. Cohen (1987, zitiert nach Marsh, 2007, S. 339) berichtete, dass die Korrelationen höher ausfielen, wenn die Studierendenleistung mit spezifischen Lehrevaluationskomponenten anhand von Multi-Item-Skalen gemessen wurde.
4. Laut Frey (1978, zitiert nach Marsh, 2007, S. 339) sind globale Ratings anfällig für Verzerrungen durch den Kontext, die Stimmung und andere Faktoren und sollten daher nicht verwendet werden. Dies sei bei Items, die das aktuelle Lehrverhalten abfragten, nicht der Fall.
5. Die Verwendung gewichteter Faktoren sei ein Kompromiss zwischen beiden Positionen.

4.2 Ablauf einer Evaluation

Idealerweise ist eine Evaluationsmaßnahme in einen systematischen Ablauf eingebettet. Hierdurch kann gewährleistet werden, dass eine Qualitätsbeurteilung transparent nach bestimmten Kriterien abläuft und festgelegt wird, bei welchem Evaluationsergebnis welche Maßnahmen zu treffen sind.

4.2.1 Ein Evaluationsprozess

Köller (2009) hat einen von Abs, Merki und Klieme (2006) skizzierten Evaluationsprozess modifiziert, der wie folgt aufgebaut ist: Demnach lässt sich eine wissenschaftlich durchgeführte Evaluation in acht Schritten in Form eines Kreislaufs unterteilen (siehe Abbildung 3), die wiederum in drei Zusammenhängen (Entstehungs-, Begründungs- und Verwertungszusammenhang) betrachtet werden können. Diese Zusammenhänge und die dazugehörigen Schritte werden im Folgenden kurz zusammengefasst und dann an einem konkreten Fall studentischer Lehrevaluation dargestellt:

1. *Entstehungszusammenhang von Evaluationen (Schritte 1 und 2):* Hierbei soll ein theoretischer beziehungsweise konzeptueller Überbau erarbeitet werden. In diesem wird klargestellt, welche Zieldimensionen aufgrund einer Maßnahme optimiert werden sollen, in welchem Kontext eine Evaluation stattfindet, welche Zielgruppe anvisiert und welches Evaluationsmodell verwendet wird.
2. *Begründungszusammenhang von Evaluationen (Schritte 3 bis 5):* In diesem Kontext sollen Fragestellungen beziehungsweise Hypothesen aufgestellt werden. Diese beschreiben, welche Maßnahmen auf welche Indikatoren in welcher Intensität und unter welchen Bedingungen wirken. Hierbei ist auch der Einbezug beziehungsweise die Konstruktion entsprechender Messinstrumente mitsamt der Testung ihrer psychometrischen Eigenschaften erforderlich. Im Anschluss daran wird die Messung durchgeführt.
3. *Verwertungszusammenhang von Evaluationen (Schritte 6 bis 8):* Hier sollen die Ergebnisse interpretiert und Konsequenzen abgeleitet werden. Dabei wird klargestellt, welche konkreten Veränderungen bei welchem Ergebnis notwendig sind. Ebenso soll auch entschieden werden, wer Zugang zu den Ergebnissen erhält.

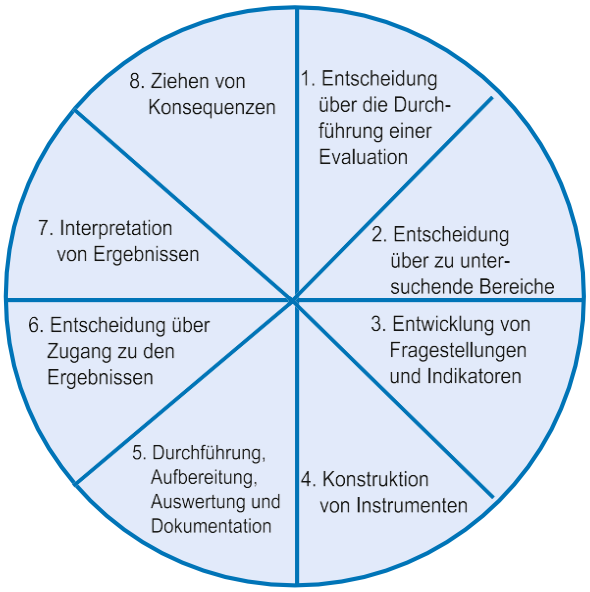

Abbildung 3: Die Schritte eines Evaluations-prozesses von Köller (2009, S. 337) modifiziert nach Abs et al. (2006)

4.2.2 Beispiel für einen Lehrevaluationsprozess

Bezogen auf das *Frankfurter Promotionskolleg am Fachbereich Medizin* sieht dieser Evaluationsprozess in folgender Weise aus (Beschreibung des Promotionskollegs siehe Kapitel 7.4):

1. Das *Frankfurter Promotionskolleg am Fachbereich Medizin* hat zum Ziel, Promovierende zu unterstützen und ihnen grundlegende Kompetenzen des wissenschaftlichen Arbeitens zu vermitteln. Dabei soll die Qualität der Lehre in dem Sinne erfasst beziehungsweise evaluiert werden, ob die Teilnehmer mit der Durchführung einer Lehrveranstaltung und der entsprechenden Vermittlung von Lehrinhalten zufrieden waren. Auf Basis dieser Erhebungen soll die Qualität der Lehrtätigkeit der Dozenten an diese rückgemeldet und bei Bedarf verbessert, die Inhalte von Lehrveranstaltungen modifiziert oder neue Angebote geschaffen werden. Diese Lehrveranstaltungsevaluationen werden nach jeder einzelnen Veranstaltung erhoben und somit wird das Promotionskolleg formativ evaluiert.
2. Die konkrete Fragestellung der Evaluation des Promotionskollegs bezieht sich darauf, ob und in welcher Form Verbesserungsbedarf hinsichtlich der Durchführung der einzelnen Lehreveranstaltung und der Vermittlung von Lehrinhalten besteht. Indikator dieses Verbesserungsbedarfs ist die Zufriedenheit der Teilnehmer, gemessen anhand verschiedener qualitätsrelevanter Kriterien, die aus sachlogischen Überlegungen, Literatur und anderen Inventaren herangezogen wurden (siehe Kapitel 8). Diese werden anhand eines Lehrevaluationsinventars erfasst (siehe Anhang A), das spezifisch für das Promotionskolleg entwickelt wurde. Nach jeder Lehrveranstaltung wird dieses Inventar an alle Teilnehmer ausgehändigt, dann eingesammelt, in eine Datenbank eingeben und ausgewertet.
3. Die Ergebnisse der Evaluation soll der Promotionskollegsleitung, dem Dekanat des Fachbereichs und insbesondere den Dozenten des Promotionskollegs zur Verfügung stehen: Je nach Ergebnis können die Dozenten ihr Lehrverhalten verändern (zum Beispiel aufgrund niedriger Werte in entsprechenden Items). Die Leitung des Promotionskollegs kann

bei durchweg negativen Evaluationen bestimmte Dozenten nicht mehr einladen und bei deutlich angemeldetem Bedarf neue Veranstaltungen etablieren. Das Dekanat kann auf Basis der Evaluationen grundlegende Entscheidungen wie zur weiterführenden Finanzierung und ähnlichem treffen.

4.3 Lehrevaluation in dieser Arbeit

In dieser Arbeit wird der Definition von Rossi und Freeman (1993, zitiert nach Bortz & Döring, 2006, S. 96) gefolgt, nach der Evaluationsforschung die systematische Anwendung empirischer Forschungsmethoden zur Bewertung des Konzeptes, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme ist. Das Interventionsprogramm ist in dieser Arbeit die universitäre Lehre, die hinsichtlich der Qualität bewertet werden soll - konkret die Qualität einer Lehrveranstaltung. Da die Zufriedenheit aus Sicht der Studierenden erfasst werden soll, findet die Bewertung anhand studentischer Lehrevaluationsinventare statt. Wegen der weiten Verbreitung des Begriffs der Lehrevaluation wird dieser beibehalten und nicht der der Lehrevaluationsforschung verwendet.

Studentische Lehrevaluationen werden häufig eingesetzt und sind in einigen Bundesländern gesetzlich vorgeschrieben (siehe Kapitel 1). Sie sind als Evaluationsart sinnvoll, da die Studierenden die direkten Adressaten von Lehre sind. Insofern ist ihre Einschätzung wichtig, ob auf ihren Bedarf hinsichtlich der Vermittlung von Lerninhalten und der Art der Durchführung einer Lehrveranstaltung angemessen eingegangen und somit eine Form von Lehrqualität erfüllt wird. Diese Beurteilung findet in dem in dieser Arbeit betrachteten Frankfurter Promotionskolleg in fortlaufender Form statt, so dass eine formative Evaluationsform vorliegt.

In dieser Arbeit wird davon ausgegangen, dass die Beurteilung auf Einzelitem-Ebene sinnvoll ist: Hinsichtlich der Veranstaltungen des Promotionskollegs ist es wichtig zu wissen, wie die einzelnen Aspekte bewertet wurden, die die jeweilige Lehrveranstaltungsdurchführung ausmachen, um sie gegebenenfalls zu verbessern (zum Beispiel, ob der Zeitrahmen angemessen war). Ein globales Rating kann als Ergänzung dienen.

Es ist allerdings zu bedenken, dass – wie weiter oben beschrieben – studentische Lehrevaluationsergebnisse relevante Konsequenzen für die individuelle Gestaltung einer Lehrveranstaltung haben können, aber auch Entscheidungen einer Universitätsleitung oder eines Fachbereichs beeinflussen können. Dementsprechend ist es wichtig, dass die Grundlage dieser Entscheidungen in Form der Lehrevaluationsergebnisse wissenschaftlichen Kriterien entspricht: Die Überprüfung dieser wissenschaftlichen Anforderung an Tests und Fragebögen wird anhand verschiedener Gütekriterien wie der *Validität*, *Reliabilität* oder *Objektivität* durchgeführt. Im Laufe der Zeit wurden viele Studien veröffentlicht, die diese Kriterien im Kontext studentischer Lehrevaluationen behandeln. Im folgenden Abschnitt werden daher die wissenschaftlichen Grundlagen erörtert, die diesen Studien zugrunde liegen und hinsichtlich ihrer Angemessenheit diskutiert.

5. Herkömmliche Vorgehensweisen bei der Überprüfung der Güte studentischer Lehrevaluationsergebnisse

5.1. Gütekriterien der Test- und Fragebogenforschung

Validität, *Reliabilität* und *Objektivität* gelten als die drei *Hauptgütekriterien* der Fragebogenkonstruktion und Testtheorie (Moosbrugger & Kelava, 2007).

5.1.1 Was bedeutet Validität?

In der Vergangenheit wurde *Validität* häufig derart definiert, dass sie vorliege, wenn ein Test das misst, was er beansprucht zu messen (Kline, 1986, S. 4), beziehungsweise, dass Validität sich damit beschäftige, was ein Test misst und wie gut er dies tut (Anastasi, 1984, S. 99).

Eine aktuelle Definition präzisiert die Bedeutung von Validität in einer komplexeren Weise: Laut den *Standards for Educational and Psychological Testing* von 2014 wird Validität in folgender Weise definiert: „Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests". Dementsprechend beinhalte der Validierungsprozess die Akkumulation relevanter Evidenz, um eine gut fundierte wissenschaftliche Grundlage für die angenommenen Testwertinterpretationen bereit zu stellen. (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014, S.11)

Neben solch allgemeinen Definitionen wurden im historischen Verlauf verschiedene Arten von Validität entwickelt: Von diesen wurden und werden häufig Kriteriums-, Inhalts-, faktorielle, Konstrukt- und Augenscheinvalidität verwandt. (siehe Fisseni, 1997, S. 95; Lienert & Raatz, 1998, S. 10 & Kap. 11)

5.1.2 Was bedeuten Reliabilität und Objektivität?

Als weiteres Hauptgütekriterium wird *Reliabilität* aufgeführt. Reliabilität beschreibt die Messgenauigkeit beziehungsweise Messfehlerfreiheit eines Messinstruments: Unter Annahme der Unabhängigkeit der Messungen voneinander wird quantifiziert, inwieweit zwei Terme zweier äquivalenter Formen eines Tests miteinander korrelieren (AERA et al., 2014, S. 33). Aspekte der Reliabilität werden häufig in der *Klassischen Testtheorie* behandelt, die dementsprechend auch Messfehlertheorie genannt wird (Moosbrugger, 2007).

Das dritte Hauptgütekriterium der *Objektivität* charakterisiert die Unabhängigkeit der Messungen von der Art der Testdurchführung, Testauswertung und Testinterpretation (Moosbrugger & Kelava, 2007, S. 8).

5.1.3 Validität und Reliabilität in der Forschung zu studentischer Lehrevaluation

Wie beschrieben können Lehrevaluationsergebnisse relevante Folgen haben. Dementsprechend wichtig ist der Nachweis von Validität hinsichtlich der Deutung und Verwendung der Ergebnisse.

In der Forschung zu studentischen Lehrevaluationsinventaren werden regelmäßig Aspekte der Validität und der Reliabilität wissenschaftlich untersucht und diskutiert (siehe Kapitel 5.6): Dazu zählen insbesondere die verschiedenen Arten von Validität (vornehmlich kriteriumsbezogene, inhaltsbezogene und faktorielle), Aspekte der Reliabilität (Klassischen Testtheorie und Generalisierbarkeitstheorie) und auch der Item Response-Theorie.

Diese werden im weiteren Verlauf dieses Kapitels erläutert, hinsichtlich ihrer theoretischen Hintergründe eingeordnet und in Bezug zu studentischen Lehrevaluationsinventaren gesetzt. Auf Objektivität wird aus Gründen der Fokussierung auf die Diskussion von Validität und Reliabilität im Kontext studentischer Lehrevaluationen in dieser Arbeit nicht eingegangen.

5.2 Der „klassische Validitätsansatz“

Mit dem Begriff des *klassischen Validitätsansatzes* werden in dieser Arbeit zwei Annahmen zusammengefasst: Zum einen die Perspektive, dass es verschiedene Arten von Validität gebe und zum anderen die Perspektive, dass ein Test an sich valide sein kann beziehungsweise zu validieren ist (Diskussion dieser Perspektiven in Abschnitt 6.1). Dementsprechend wird ein Test unter anderem als kriteriumsvalide, inhaltsvalide, eindimensional beschrieben (Überblick zu Validitätsarten zum Beispiel bei Fisseni, 1997, S. 95; Lienert & Raatz, 1998, S. 10 & Kap. 11).

Im Folgenden werden Validitätsarten vorgestellt, die häufig bei der wissenschaftlichen Begutachtung und Testung von Lehrevaluationsinventaren verwendet werden.

5.2.1 Kriteriumsvalidität

Beschreibung

Kriteriumsbezogene Validität beschreibt die Effektivität eines Tests, ein individuelles Verhalten in einer bestimmten Situation vorherzusagen. Dabei wird jeweils das Testverhalten anhand eines Korrelationskoeffizienten mit einem Kriterium verglichen. Dieses Kriterium soll ein direktes und unabhängiges Maß von dem Merkmal sein, für dessen Vorhersage der Test konstruiert wurde. (Anastasi, 1984, S. 105)

Der Nachweis von kriteriumsbezogener Validität kann prognostisch (mit einem in der Zukunft liegenden Kriterium), konkurrent (mit einem Kriterium, das zeitgleich erfasst wird) und retrograd (mit einem Kriterium, das in der Vergangenheit liegt) durchgeführt werden (Rammstedt, 2004, S. 17).

Beispiel

Eine Studie zu einem Instrument zur Messung von *Selbstkontrolle* untersuchte Validität unter anderem anhand der Übereinstimmung mit Kriterien. Selbstkontrolle wurde als die „Überwindung oder Modifikation von Reaktionstendenzen“ definiert und anhand der deutschen Version der *Self-Control Scale* gemessen. Für die Selbstkontroll-Kapazität der Regulierung von Impulsen, Gedanken, Affekten und leistungsbezogenem Verhalten wurden Zusammenhänge mit Kriterien, die diesen Bereichen zugeordnet werden können, vermutet und bestätigt. Dies waren zum Beispiel negative, mittel-hohe Korrelationen hinsichtlich der Neigung zu *impulsivem Aggressionsausdruck* und zu *Prokrastination*. (Bertrams & Dickhäuser, 2009)

5.2.2 Inhaltsvalidität

Beschreibung

In der Ausgabe der *Standards for Educational and Psychological Testing* von 1974 wurde die Inhaltsvalidität eines Tests als gegeben angenommen, wenn seine Items eine repräsentative Stichprobe aus einer Itemmenge darstellten (APA, 1974). Klauer (1984, S. 1) legte eine ausführlichere Definition vor: Ein Test sei inhaltsvalide, wenn die zu ihm gehörenden Items eine repräsentative Stichprobe für eine zuvor definierte Grundgesamtheit darstellten. Diese Itemmenge sei eindeutig definiert, wenn der fragliche Sachverhalt klar dargestellt, eine Itemform gewählt wurde und wenn Transformationsregeln beschrieben seien, wie der Sachverhalt in Items abzubilden ist.

Eine Vorgehensweise zur Bestimmung der Inhaltsvalidität kann beispielsweise durch Hinzunahme kompetenter Beurteiler erfolgen, die nach einem Ratingverfahren über den Grad der Validität eines Tests befragt werden (Lienert & Raatz, 1998).

Als Teilbereiche der Inhaltsvalidität werden *Sampling*- und *Item*-Validität beschrieben (zum Beispiel bei Lodico, Spaulding & Voegtle, 2006 oder Spooren, Brockx & Mortelmans, 2013): Sampling-Validität sei auf die Repräsentativität aller Items bezüglich des inhaltlichen Bereiches bezogen, beziehungsweise darauf, ob das Instrument somit als Ganzes den interessierenden Bereich repräsentiere. Item-Validität beziehe sich darauf, ob jedes einzelne Item ein Maß des gewünschten inhaltlichen Bereiches ist.

Beispiel

An einem Test zur *Modellkompetenz im Kontext von Biologieunterricht* (Terzer, Hartig & Upmeier zu Belzen, 2013) soll eine Testkonstruktion in sieben Schritten dargestellt werden, die Inhaltsvalidität gewährleisten soll:

1. *Zunächst definierten die Autoren den Untersuchungsgegenstand der Modellkompetenz im Kontext von Biologieunterricht und dessen theoretische Fundierung:*
    - Was wird unter *Kompetenz* verstanden? Verfügbare Fähigkeiten und Fertigkeiten.
    - Welche Aspekte von Kompetenz interessieren? Nur die kognitiven Aspekte hinsichtlich der Anwendung von Modellen, keine motivationalen.
    - Was wird unter *Modell* verstanden? Bedingungen, unter denen ein Subjekt einen Gegenstand als Modell versteht und entsprechend nutzt, und es somit als Methode im Sinne einer naturwissenschaftlichen Arbeitsweise diene.
    - Die theoretische Fundierung des Untersuchungsgegenstandes basiert auf Grundlage der Zielgruppe (Schüler der siebten bis zehnten Jahrgangsstufe des Gymnasiums und der Realschule), und es wurden Ergebnisse empirischer Studien an Schülern, Lehrern und Experten herangezogen.

- Arbeitsdefinition des Konstrukts als Grundlage für die Operationalisierung der Testitems: Ein *Kompetenzmodell* untergliedert in fünf Teilkompetenzen mit je drei Niveaus und in zwei Dimensionen gruppiert (Kenntnisse über Modelle und Modellbildung).

2. *Die Testkonzeption:*
    - Klärung des Ziels der Untersuchung: Die empirische Überprüfung eines Kompetenzmodells an Schülern und deren Verständnis konkreter Modelle und deren konkreter Umgang mit Modellen der Biologie.
    - Die Art der Messung in Anbetracht des spezifischen Kontexts: Leistungstests mit Items für alle Aspekte des Modells.
3. *Systematisierung der Itemkonstruktion:* Formulierung von Indikatoren, die sich jeweils auf einen Bereich der Kompetenz beziehen. Die Beschreibung des Itemuniversums und die Wahl der Antwortformate.
4. *Entwicklung einer Konstruktionsanleitung:* Experten aus der empirischen Bildungsforschung prüften, inwiefern die Konstruktionsanleitung das zu Grunde liegende Modell angemessen operationalisiert.

5. bis 7.: In diesen Schritten werden die Items entwickelt, erprobt, selektiert und das Erhebungsdesign festgelegt.

5.2.3 Faktorielle Validität

Beschreibung

Nach Piedmont (2014, S. 2148) wird unter faktorieller Validität das Ausmaß verstanden, in dem die vermeintlich einer Skala zu Grunde liegende Struktur (bestehend aus Konstrukten und Items) in einer bestimmten Konstellation reproduzierbar ist. Dabei ist anzugeben, was die entsprechenden Konstrukte ausmacht, und durch welche Items diese definiert werden. Diese Strukturen können in folgenden Varianten auftreten:

- eindimensional: Items hängen in einer Dimension zusammen.
- mehrdimensional: Mehrere voneinander unabhängige Dimensionen, bei denen die Items über die Dimensionen hinweg nicht zu einem Faktor zusammengefasst werden können.

- mehreren miteinander korrelierenden Facetten mit jeweils eigener Interpretation, aber die Items können in einer kohärenten Dimension aggregiert werden (multifaceted).

In Faktorenanalysen ergibt sich ein empirisch beobachteter Wert *x* einer Person *p* auf einem Item *i* aus dem mit einer entsprechenden Faktorladung λ gewichteten Konstrukt ξ (beispielweise einer Persönlichkeitseigenschaft einer Person) und dem Residuum δ (siehe Gleichung 1, die sich auf konfirmatorische Faktorenanalysen bezieht).

$$x_{pi} = \lambda_i \xi_p + \delta_{pi} \qquad (1)$$

Überprüft wird diese Validitätsart anhand *konfirmatorischer* (Brown, 2006) oder *exploratorische Faktorenanalysen* (Fabrigar & Wegener, 2012): Konfirmatorische Faktorenanalysen sind derart, dass vorab theoriegeleitet festgelegt wird, welche Items auf welche Konstrukte rückführbar sind. Somit können auch konkurrierende Theorien an einem Datensatz gegeneinander getestet werden. Exploratorische Faktorenanalysen sollen die Anzahl unterschiedlicher Konstrukte ermitteln, um das Korrelationsmuster einer Ansammlung von Messungen zu erklären. Die Ergebnisse werden dann im Nachhinein interpretiert. Die Items eines Tests werden in den Analysen als *manifeste* und die Konstrukte als *latente Variablen* spezifiziert.

Faktorenanalysen werden auch als Erweiterung der *Klassischen Testtheorie* (siehe Abschnitt 5.3) angesehen, denn mit der Schätzung des Konstrukts auf latenter Ebene wird versucht, den Messfehler auf manifester Ebene zu kontrollieren (Rauch & Moosbrugger, 2011). Weiterhin gilt sie als Voraussetzung zur Prüfung von Konstruktvalidität (Cronbach & Meehl, 1955).

Beispiel

Ursprünglich wurden Faktorenanalysen insbesondere in der Persönlichkeitsforschung eingesetzt: Beispielsweise wurde das Antwortverhalten bei Intelligenztests dahingehend faktorenanalytisch untersucht, inwiefern Items gemeinsam variieren, um somit herauszufinden, aus welchen Domänen Intelligenz besteht und welche Items als Indikatoren dieser Domänen anzusehen sind. Im *Primärfaktorenmodell* der Intelligenz von Thurstone (1938; zitiert nach Stemmler, Bartussek, Hagemann & Amelang, 2011, S. 149) ist die Leistung in einer

Satzergänzungsaufgabe als Ausdruck des Konstrukts *Sprachverständnis* anzusehen und Multiplikationsaufgaben als Ausdruck des Konstrukts *Rechenfähigkeit* (number).

Ebenso wurden Persönlichkeitsdimensionen wie die *Big Five* auf Basis vieler Fragebogenitems anhand von Faktorenanalysen extrahiert (MacCrae, 2009; zitiert nach Myers, 2014, S. 573). Nach diesem Modell sind sämtliche Persönlichkeitsaspekte Ausdruck von fünf voneinander unabhängigen Faktoren beziehungsweise Dimensionen (siehe Abbildung 4).

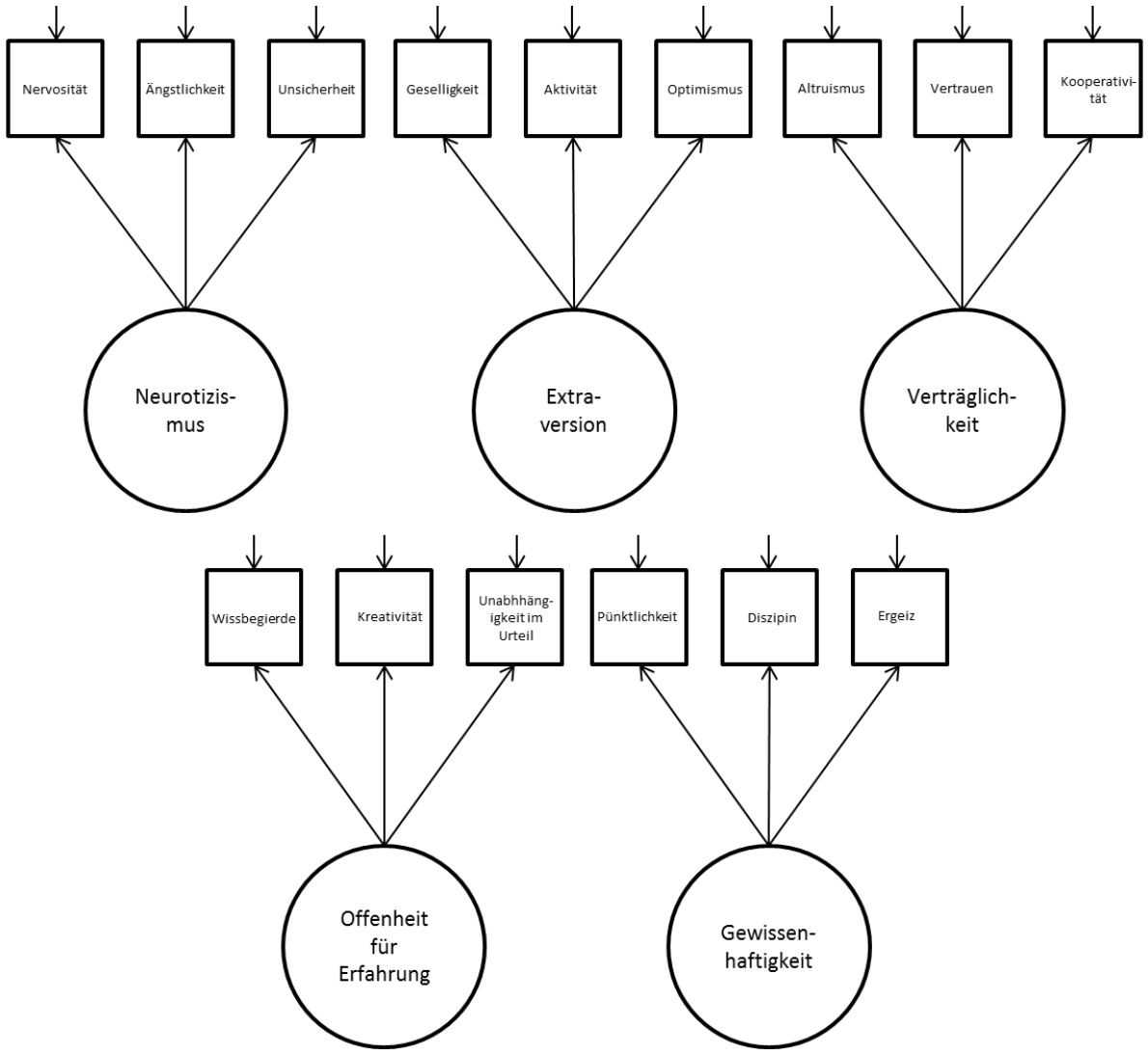

Abbildung 4: Beispielhafte Skizzierung der Faktorstruktur der Big Five

5.3 Klassische Testtheorie

5.3.1 Beschreibung

Die *Klassische Testtheorie* (KTT) ist maßgeblich ein Ansatz zur Bestimmung der Messgenauigkeit (Reliabilität) beziehungsweise des Messfehlers eines Tests zur Erfassung von Merkmalsausprägungen. Dementsprechend wird sie auch Messfehlertheorie genannt. Die KTT behandelt somit nicht Fragen der Validität. Allerdings gilt gegebene Reliabilität als eine ihrer Voraussetzungen, da eine ungenaue Messung keine angemessenen Rückschlüsse hinsichtlich der Ausprägungen von Merkmalen zulässt. (Moosbrugger, 2007)

Wie in Gleichung 2 dargestellt, besteht nach der KTT ein Testwert x_{pi} aus einem sogenannten wahren Wert τ_{pi} und einem unsystematischen Messfehler ε_{pi}, wobei die Messfehler zwischen den jeweiligen Items *i* und Personen *p* sowie mit den wahren Werten als unabhängig voneinander angenommen werden (Moosbrugger, 2007).

$$x_{pi} = \tau_{pi} + \varepsilon_{pi} \qquad (2)$$

Der Anteil des Messfehlers kann anhand verschiedener Methoden bestimmt werden: Die *Retest-Reliabilität* bestimmt – bei Annahme der zeitlichen Konstanz der Ausprägung des zu messenden Merkmals – die Messfehlerfreiheit über mehrere Messzeitpunkte. Wenn ein Test nur zu einem Zeitpunkt erhoben wird, kann – unter Annahme von Itemhomogenität - die Übereinstimmung von zwei unterschiedlichen Varianten eines Tests mit gleicher Messgenauigkeit (*Paralleltest-Reliabilität*), zwei Testhälften (*Split Half-Reliabilität*) oder zwischen jedem einzelnen Testitem, das als jeweils ein eigenständiger Test angesehen wird, bestimmt werden (*interne Konsistenz* wie zum Beispiel *Cronbach* α). In Gleichung 3 wird die Formel von Cronbach α dargestellt: Hierbei wird die Summe der Varianz der einzelnen Testitems $Var(x_i)$ an der Varianz des Gesamttests $Var(x)$ relativiert. Durch *m* wird die Anzahl der Items des jeweiligen Tests in die Formel eingebracht. (Schermelleh-Engel & Werner, 2007)

$$Rel(x) = a = \frac{m}{m-1} \cdot \left\{1 - \frac{\sum_{i=1}^{m} Var\,(x_i)}{Var\,(x)}\right\} \qquad (3)$$

Das Ergebnis repräsentiert in Form eines Reliabilitätskoeffizienten das Verhältnis zwischen wahrer und Gesamtvarianz und nimmt einen Wert zwischen *0* und *1* an. Je höher der Wert des Koeffizienten, als desto niedriger wird der Messfehler

angenommen und je niedriger der Wert, desto höher der Messfehler. Eine niedrige Reliabilität zeigt an, dass die verschiedenen Items zu einem hohen Anteil nicht gemeinsam miteinander variieren sondern durch unsystematische Messefehler beeinflusst wurden. Somit wären nur ungenaue oder keine Aussagen über individuelle Testwerte und keine messgenaue Differenzierung zwischen Testpersonen möglich. Als Mindesthöhe wird beispielsweise nach Fisseni (1997, S. 124) eine Reliabilität von .80 vorgeschlagen. Bühner (2012, S. 139) merkt dazu an, dass solche Beurteilungen der Reliabilitätshöhe vom Kontext abhingen, wie der Art und Breite des untersuchten Merkmals und der Homo- oder Heterogenität der Stichprobe.

Eine bedeutsame Voraussetzung der Kennwerte der Klassischen Testtheorie ist die bereits genannte Itemhomogenität: Diese beschreibt die Eigenschaft der Items, jeweils dasselbe Merkmal zu messen. Diese wurde und wird teilweise auch heute noch durch hohe Item-Interkorrelationen, Item-Trennschärfen oder interne Konsistenzen als gegeben angesehen. Allerdings haben Studien gezeigt, dass ein Nachweis nicht ausreichend durch diese Parameter gewährleistet werden kann, sondern durch Verfahren wie exploratorische und konfirmatorische Faktorenanalysen angemessen überprüft werden sollte (Green, Lissitz & Mulaik, 1977; Gerbing & Anderson, 1988).

Sich bei der Bestimmung der Itemhomogenität nicht nach einem beispielsweise hohem Cronbach α zu orientieren ist bedeutsam: Auf Basis der Spearman-Brown-Formel (Schermelleh-Engel & Werner, 2007) kann errechnet werden, in wieweit durch eine Erhöhung der Itemanzahl die Reliabilität angehoben oder bei einer ausreichend hohen internen Konsistenz die Skala aus Gründen der Testökonomie gekürzt werden kann. Bei gegebener Itemhomogenität führt eine Verdoppelung der Testlänge zu einer Vervierfachung der wahren Varianz bei Verdoppelung der Fehlervarianz. Aber auch wenn Itemhomogenität nicht vorliegt, kommt es bei einer Testverlängerung auf Basis der Spearman-Brown-Formel zu einer Erhöhung des Reliabiliätskoeffizienten, die gegebenenfalls fälschlich als gemeinsame Merkmalsvarianz gedeutet würde.

Ein weiterer Aspekt, der der KTT zugeordnet wird, ist die *Itemanalyse*: Diese beinhaltet die deskriptivstatistische Evaluation jeden einzelnen Items:

- *Itemschwierigkeit:* Entspricht dem prozentualen Anteil einer Stichprobe, nach dem eine Aufgabe richtig beziehungsweise eine Frage oder Aussage positiv

beantwortet wurde. Je größer der daraus berechnete Index, desto leichter ist das Item zu beantworten beziehungsweise dem Inhalt zuzustimmen.

- *Itemvarianz:* Maß der Differenzierungsfähigkeit eines Items.
- *Korrigierte Itemtrennschärfe:* Gibt anhand eines Korrelationsparameters an, wie stark die Differenzierung eines Merkmals anhand eines einzelnen Items mit der des restlichen Gesamttests übereinstimmt.

Items mit einem zu niedrigen oder zu hohen Trennschärfe- oder Schwierigkeitsparameter können aus einer Skala entfernt werden: Bei der Trennschärfe differenzieren sie dann entweder zu gering im Vergleich mit der Skala oder sind aufgrund einer zu hohen Übereinstimmung überflüssig. Bei der Schwierigkeit sind sie entweder zu leicht zu oder zu schwer für eine bestimmte Population zu beantworten.

5.3.2 Beispiel

Ein Fragebogen zur Erfassung *Volitionaler Komponenten im Sport (VKS)* (Komponenten wie *Beharrlichkeit* und *Fähigkeit zur Selbstregulation*) wurde anhand mehrerer Stichproben für eine Validierungsstudie untersucht. Zunächst wurden anhand einer Itemanalyse mit Kennwerten zu Itemschwierigkeit und Trennschärfe sowie inhaltlicher Redundanz 32 von 92 Items entfernt. Eine darauffolgende Hautpkomponentenanalyse ergab vier Faktoren (*Selbstoptimierung*, *Aktivierungsmangel*, *Fokusverlust* und *Selbstblockierung*), die gering bis mittel miteinander korrelierten. Die Subskalen wiesen interne Konsistenzen (Chronbach α) von .76 bis .92 auf. Dieses Ergebnis führte die Autoren zu der Schlussfolgerung, dass der Fragebogen ein reliables Verfahren sei. (Wenhold, Elbe & Beckmann, 2009)

5.4 Generalisierbarkeitstheorie

5.4.1 Beschreibung

Die *Generalisierbarkeitstheorie* (GT) behandelt die Frage nach der *Generalisierbarkeit* einer Messung über verschiedene Bedingungen und Kontexte hinweg (Eisend, 2007) und bewertet somit die *Zuverlässigkeit* von Messungen (Cronbach, 1972).

Diese Zuverlässigkeit bezieht sich auf die Genauigkeit der Generalisierbarkeit eines beobachteten Testwertes einer Person in Bezug zu dem mittleren Wert, den diese Person nach Teilnahme unter allen möglichen Bedingungen erhalten würde. Die

GT setzt als Annahme voraus, dass das gemessene Merkmal stabil ist und Testwert-Unterschiede eines Individuums in verschiedenen Messsituationen von einem oder mehreren Fehlerquellen abhängig sind und nicht systematischen Veränderungen aufgrund von Reife oder Lernen unterliegen. (Shavelson & Webb, 1991)

Die GT wird als Erweiterung der Klassischen Testtheorie (KTT) angesehen und lässt somit auch Aussagen zur Messfehlerfreiheit zu: Die Messfehlerbestimmung bei der KTT wird bei der Messung von Unterschieden zwischen Personen eingesetzt, kann aber nicht verschiedene Messfehlerquellen erfassen. In der KTT wird zwischen *wahrem Wert* und *unsystematischen Messfehler* unterschieden. In der GT wird der Begriff des *wahren Werts* durch den des *globalen wahren Wertes (universe Score)* ersetzt. Dieser Wert ist der geschätzte Mittelwert über alle Bedingungen hinweg und entspricht dem Erwartungswert einer Messung. Hinsichtlich des Messfehlerterms werden in der GT zusätzlich weitere systematische Varianzquellen beziehungsweise systematische Messfehler (sogenannte Facetten) angenommen, die einen empirischen Testwert beeinflussen. Der Messfehler wird daher theoriegeleitet in mehrere Facetten zerlegt, und diese werden anhand einer Varianzkomponentenschätzung analytisch überprüft und quantifiziert. Durch eine Varianzkomponentenschätzung kann verstanden werden, wie unerwünschte Varianz entsteht, und ein effizientes Design für die Sammlung weiterer Daten entworfen werden. (Cronbach, 1972; Eisend, 2007 oder Rauch & Moosbrugger, 2011)

In Gleichung 4 wird dargestellt, wie sich der empirisch beobachtete Wert X beispielsweise einer Person p bei einem Item i durch den Gesamtmittelwert (grand mean) μ und dem Einfluss der einzelnen Person $(\mu_p - \mu)$ und dem Einfluss des einzelnen Items $(\mu_i - \mu)$ und einem möglichen $(X_{pi} - \mu_p - \mu_i + \mu)$ Interaktionseffekt inklusive Residuum zusammensetzt.

$$X_{pi} = \mu + \mu_p - \mu + \mu_i - \mu + X_{pi} - \mu_p - \mu_i + \mu \qquad (4)$$

Nach Gleichung 5 setzt sich die Varianz der Personen p auf den Items i aus den jeweiligen Varianzkomponenten und dem Residuum inklusiver potentieller Interkationen zusammen.

$$\sigma^2(X_{pi}) = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 \qquad (5)$$

Ein auf Basis der Schätzung errechneter *Generalisierbarkeitskoeffizient* (analog zum Reliabilitätskoeffizienten der KTT) gibt an, wie präzise die Generalisierung eines beobachteten Testwerts einer Person - auf Basis einer Verhaltensstichprobe dieser Person - hinsichtlich ihres globalen wahren Wertes ist. Dabei wird dessen Varianz an der Summe der Varianzkomponenten relativiert. Je höher die Ausprägungen des Koeffizienten, desto eher kann ein Testwert als konstant über verschiedene Bedingungen hinweg angesehen werden. Er kann in Form zweier Varianten berechnet werden: Für relative Entscheidungen (hier zählen Varianzkomponenten zur Fehlervarianz, die den relativen Platz eines Untersuchungsobjekts beeinflussen) und absolute Entscheidungen (diejenigen Varianzkomponenten zählen zur Fehlervarianz, die den absoluten Platz eines Untersuchungsobjekts beeinflussen; unabhängig von anderen Untersuchungsobjekten).

In Gleichung 6 wird die Berechnung des Generalisierbarkeitskoeffizienten ϕ für absolute Entscheidungen dargestellt: Die Varianz des globalen wahren Werts σ_p^2 - bei Spezifikation der Personen p als Facette der Diskriminierung - wird an der Gesamtvarianz relativiert. Diese besteht aus der Varianz des globalen wahren Werts σ_p^2 und der absoluten Fehlervarianz σ_{abs}^2.

$$\phi = \left(\frac{\sigma_p^2}{\sigma_p^2 + \sigma_{abs}^2}\right) \qquad (6)$$

Das analytische Vorgehen auf Basis der Generalisierbarkeitstheorie wird in zwei Schritte unterteilt:

1. *Die Generalisierungs-Studie (G-Studie):* In der G-Studie werden die Facetten der Diskriminierung (erwünschter Varianzeinfluss) und die Facetten der Generalisierung (unerwünschte Varianzquellen beziehungsweise die Bedingungen, über die ein Testwert hinweg generalisiert werden sollte) spezifiziert. Auf dieser Grundlage werden anhand von Daten die Varianzkomponenten geschätzt und der Generalisierbarkeitskoeffizient berechnet.
2. *Die Entscheidungs- beziehungsweise Decision-Studie (D-Studie):* In der gegebenenfalls auf die G-Studie folgende D-Studie kann bei einem als zu niedrig angesehenen Generalisierbarkeitskoeffizienten ein Design gestaltet werden, um die Messungen optimal zu gestalten: Zum Beispiel, um wie

viele Abstufungen die Varianzquellen erhöht werden müssen, um von einer generalisierbaren Aussage des Testwertes ausgehen zu können (analog zur Spearman-Brown-Formel in der KTT). Dies kann je nach Fall beispielsweise eine Erhöhung der Anzahl der Erhebungszeitpunkte oder der Anzahl an Ratern sein. Ebenso kann bei einer günstigen Ausgangslage eine Reduktion der genannten Bedingungen sinnvoll sein, um ein beispielsweise kostengünstigere Erhebungsdesign zu haben und trotzdem ein ausreichend generalisierbares Ergebnis zu erhalten. Die anzustrebende Höhe des Koeffizienten sollte sich an den Konsequenzen des Testeinsatzes und denen der Entscheidungsstudie orientieren. Als Daumenregel gibt Nunnally Nunnally (1978, zitiert nach Finn & Kayande, 1997, S. 264) einen anzustrebenden Wert von mindestens .90 und idealerweise ab .95 an.

5.4.2 Beispiel

Anhand einer hypothetischen Studie von Eisend (2007) kann die Generalisierbarkeitsstudie veranschaulicht werden: Hierbei wird dargestellt, dass die Prüfungsleistungen Studierender beispielsweise auch vom Prüfungszeitpunkt und der Prüfungsform beeinflusst werden können (siehe Venn-Diagramm in Abbildung 5). Somit würden die Studierenden die Facette der Diskriminierung und Prüfungsform sowie Prüfungszeitpunkt die Facetten der Generalisierung darstellen. Hierbei kann überprüft werden, ob einzelne Prüfungsleistungen der Studierenden unabhängig von Form und Zeitpunkt sind und somit über verschiedene Bedingungen generalisiert werden können. Zusätzlich können die Interaktionen der verschiedenen Facetten miteinander überprüft werden: Diese Interaktionen können aber nur bestimmt werden, wenn ein vollständiges Design vorliegt, also jeder Studierende jede Prüfungsform zu jedem Zeitpunkt mitgemacht hat (Cronbach, 1972, S. 45). Als beispielhaftes Ergebnis wurde aufgeführt, dass die Facette der Studierenden mit annähernd 26% und die der Prüfungsform mit nahezu 46% einen Großteil der Varianz erklären und der Prüfungszeitpunkt mit etwas mehr als 5% keinen bemerkenswerten Anteil besitzt. Dieses Ergebnis impliziere, „dass die geprüfte Leistung von Studierenden nicht ohne weiteres über verschiedene Prüfungsformen hinweg generalisiert werden kann“ (Eisend, 2007, S. 11).

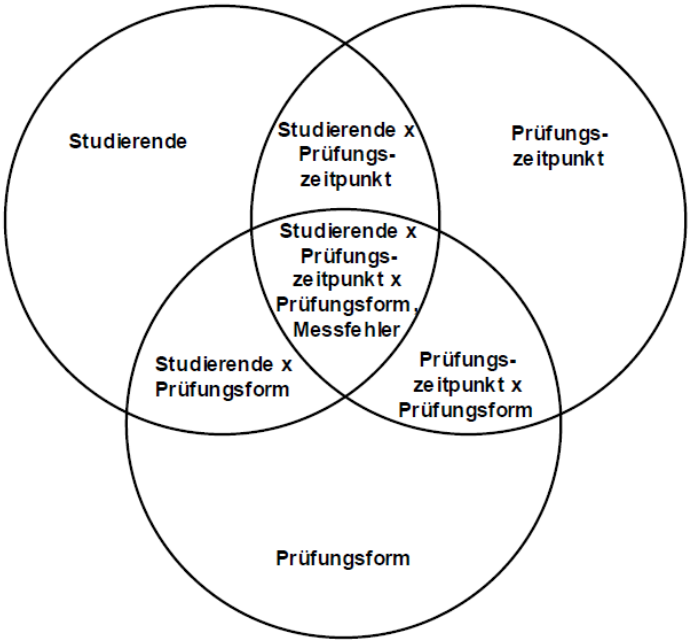

Abbildung 5: Die Darstellung der Varianzkomponenten in einem Venn-Diagramm (entnommen aus Eisend, 2007, S. 6)

5.5 Item Response-Theorie

5.5.1 Beschreibung

Eine weitere Theorie zur Überprüfung von Tests und Fragebogen ist die *Item Response-Theorie (IRT)* oder auch *Probabilistische Testtheorie* genannt. Die IRT ist eine modellbasierte Messung, in der die Schätzungen der Ausprägungen der Eigenschaften der untersuchten Personen von den Itemantworten dieser Personen als auch den Eigenschaften der Items abhängen (Embretson & Reise, 2000). Diese können gemeinsam in einer so genannten *Itemcharakteristischen Funktion (ICC)* abgebildet werden (siehe Abbildung 6).

Anhand der IRT können Fragen der Dimensionalität von Tests (ähnlich der faktoriellen Validität) und der Güte der einzelnen Items behandelt werden.

Wie bei den Faktorenanalysen werden das zu messende Merkmal als latente und seine Indikatoren als manifeste Variablen spezifiziert. Es gibt Modelle für Items mit binärem (zum Beispiel Rasch-Modell), ordinalem (zum Beispiel das Partial Credit-Modell) und metrischem Antwortformat. Auch kann die Ein- oder Mehrdimensionalität eines Tests geprüft werden.

Anhand der IRT können neben dem *Personen-* verschiedene *Item-Parameter* geschätzt werden, wobei sich die Möglichkeit zu deren Berechnung zwischen den verschiedenen Modellen unterscheidet:

- *Personenparameter:* Die Ausprägung des interessierenden Merkmals einer Person auf einer latenten Variable.
- *Schwierigkeitsparameter:* Ist durch den Wert einer latenten Merkmalsausprägung definiert, bei der die Lösungs- oder Zustimmungswahrscheinlichkeit des Items bei 50% liegt.
- *Diskriminationsparameter:* Gibt an, wie stark sich Lösungs- oder Zustimmungswahrscheinlichkeit in Abhängigkeit von der Merkmalsausprägung verändert.
- *Rateparameter:* Wird modelliert, wenn bei dem Antwortverhalten bei einem Leistungstest mit Multiple-Choice-Aufgaben die korrekte Antwort geraten werden könnte.

In Gleichung 7 wird ein bestimmtes IRT-Modell (Einparameter-Logistisches Modell aus der Gruppe der Rasch-Modelle) beispielhaft dargestellt, in dem es die Wahrscheinlichkeit *P* für eine von zwei möglichen Angaben *x* auf einem Item *i* (ob beispielsweise eine Aufgabe gelöst wurde oder nicht) berechnet wird. Dabei wird die Schwierigkeit σ eines Items *i* und der die Merkmalsausprägung ξ einer Person *p* berücksichtigt.

$$\text{Gleichung: } P(x_{pi}) = \frac{\exp(x_{pi}(\xi_p - \sigma_i))}{1+\exp(\xi_p - \sigma_i)} \qquad (7)$$

Weiterhin kann anhand der Iteminformationsfunktion angegeben werden, wie hoch der Informationsgehalt eines Items hinsichtlich der Diskrimination zwischen verschiedenen Merkmalsausprägungen ist, und je nach Ergebnis Items aus einer Skala entfernt werden (Hartig, Frey & Jude, 2007).

5.5.2 Beispiel

IRT-Modelle werden häufig in der Bildungsforschung eingesetzt. So kann eine interessierende Kompetenz anhand verschiedener Items in Form von Aufgaben gemessen werden, und die jeweilige Ausprägung der Kompetenz der beteiligten Personen sowie die Schwierigkeit der einzelnen Aufgaben geschätzt werden.

Item Response-Modelle wurden beispielsweise in Form eines Rasch-Modells in der PISA 2000-Studie eingesetzt (Lind & Knoche, 2004): Es wurden ein Personen- beziehungsweise Fähigkeitsparameter für das Kompetenzniveau der Schüler und ein Schwierigkeitsparameter für die jeweiligen Aufgaben aus den Bereichen der Lesekompetenz, Mathematik und Naturwissenschaften geschätzt. Das daraus resultierende Modell für das jeweilige Item kann in der Itemcharakteristischen Funktion abgebildet werden (siehe Abbildung 6).

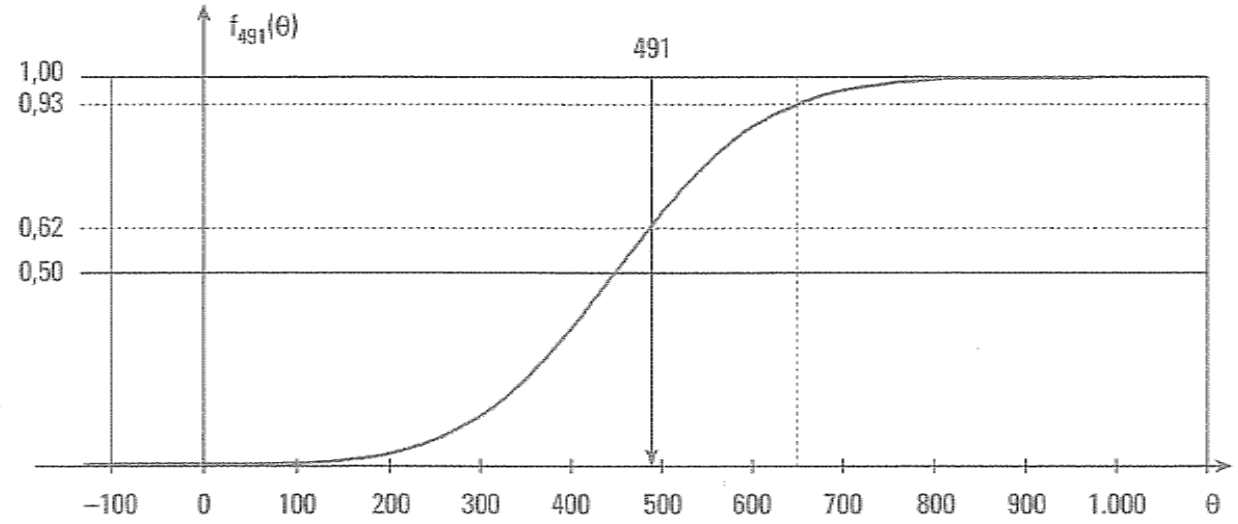

Abbildung 6: Itemcharakteristische Funktion einer Textaufgabe zur Prozentrechnung („Glasfabrik, Version 2"). 65% der Probanden konnten sie lösen, eine Person mit einem Fähigkeitswert von 491 hat eine Lösungswahrscheinlichkeit von 62%. Abbildung entnommen aus Lind & Knoche (2004, S. 61).

5.6 Die Anwendung der Ansätze bei Lehrevaluationsinventaren

In vielen wissenschaftlichen Arbeiten zur studentischen Lehrevaluation wurden die in den vorhergehenden Abschnitten aufgeführten Kriterien zur Validität und Messefehlerfreiheit untersucht. Im Folgenden werden diese jeweils beispielhaft an bereits publizierten Studien erläutert und diskutiert.

Dies wird anhand bisher entwickelter und publizierter Lehrevaluationsinventare dargestellt: Es wurden allerdings national und international eine große Menge an Inventaren entwickelt. Um einen überschaubaren Rahmen beizubehalten und aufgrund des Bezugs auf das in Deutschland gegebene universitäre System, wird sich in dieser

Arbeit großteilig auf eine Auswahl von Inventaren des deutschsprachigen Raums bezogen, an denen eine Überprüfung verschiedener Gütekriterien stattfand.

5.6.1 Kriteriumsvalidität:

Bedeutung

Laut Rindermann (2009, S. 166) ist für eine enge Validitätsbestimmung studentischer Lehrevaluationsergebnisse die Hinzunahme externer valider Kriterien notwendig. In der Literatur wurden als Kriterien *Lernerfolg* sowie *Selbst-* und *Fremdeinschätzungen* aufgeführt.

Lernerfolg als Validitätskriterium

Ein striktes externes Kriterium wäre laut Marsh (1987, S. 13) und Rindermann (2009, S .166) *Lernerfolg*, das der Leistung der Studierenden entspricht.

Mehrere Studien untersuchten den Zusammenhang von Evaluationsergebnissen und der Leistung von Studierenden: Cohen (1981) fasste anhand eines metaanalytischen Ansatzes den Zusammenhang zwischen der Leistung der Studierenden und der Beurteilung der Kurse oder des Dozenten zusammen. Die Leistung der Studierenden wurde anhand einer gewöhnlichen Abschlussprüfung, die Gesamtbeurteilung des Dozenten und der des Kurses wurden jeweils durch ein Einzelitem oder durch den Mittelwert aller den Dozenten oder den Kurs betreffenden Items gemessen. Dabei zeigte sich eine mittlere Korrelation von r = .47 zwischen der Leistung und der Kurs-Gesamtbeurteilung und eine mittlere Korrelation von r = .43 mit der Gesamtbeurteilung der Dozenten. In einer weiteren Metaanalyse mit 43 Studien zu *Multisection Studies* von d'Apollonia & Abrami (1996, zitiert nach d'Apollonia & Abrami, 1997, S. 1202) zeigte die Skala *allgemeine Lehrkompetenz* einen Zusammenhang mit *Lernerfolg* von r = .33 (bei Reliabilitätskorrektur .47).

Rindermann (2009, S. 166) merkte an, dass Prüfungen in der Regel nur Wissen geringen Elaborationsgrades erheben würden und nicht Lehrqualität, die sich auf Verständnis und Tiefenlernstrategien beziehe. Lernerfolg wäre somit als distales Kriterium anzusehen, da Eigenarbeit, Vorwissen, Intelligenz, Vorbereitungszeit zwischen Lernerfolg und Lernen stünden (Kromrey, 1993; zitiert nach Rindermann, 2009, S. 166).). In diesem Sinne wären die Förderung eigenständigen Denkens und Einstellungswandel weitere Kriterien von Lehrevaluationsergebnissen (Winteler, 1974 und Keil, 1975; zitiert nach Rindermann, 2009, S. 166).

Wenn Lehrqualität allerdings durch das Erreichen von Zielen überprüft werden sollte, dann sind die Zielinhalte mit den Ergebnissen abzugleichen, zum Beispiel in der Form, wie sich die tatsächliche Abschlussnote im Vergleich zu der in den Zielen angestrebten verhält.

Fremd- und Selbsteinschätzung als Validitätskriterien

Rindermann (2009, S. 167) schlug als weitere Kriterien den Vergleich von Selbst- und Fremdeinschätzung von Dozenten vor: Die alleinige oder vorwiegende Verwendung von Selbsteinschätzungen der Dozenten sei problematisch, da ihnen im Gegensatz zu den Studierenden der Vergleichsmaßstab fehle. Dementsprechend müssten Fremdeinschätzungen Dritter hinzugezogen werden, die beispielsweise durch interne oder externe Kollegen, Hochschuldidaktiker, Berufspraktiker oder Absolventen durchgeführt werden könnten.

In einem Review (Feldman, 1989) wurden die Gesamt-Beurteilungen (overall-ratings) verschiedener Beurteiler bezüglich der Effektivität von College-Dozenten verglichen: Derzeitige und frühere Studierende, Kollegen, Angehörige der Universitätsleitung (Administrators), externe neutrale Beobachter und die Dozenten selbst. Es wurde eine „relative Ähnlichkeit“ in Form von Korrelationen der jeweiligen Beurteilungen dargestellt. Zu 5 der 15 möglichen bivariaten Korrelationen gab es eine angemessene Datengrundlage, und es zeigten sich folgende Ergebnisse:

- Die höchste relative Ähnlichkeit (r = .55) zeigte sich beim Vergleich der Beurteilungen durch die aktuellen Studierenden mit denen der Kollegen der Dozenten.
- Studierende und Leitungsangehörige zeigten (r = .39) gewisse Übereinstimmungen.
- Kollegen und Leitungsangehörige waren relativ gleich (r = .48).
- Die Selbstbeurteilung von Dozenten und deren Beurteilung durch die Studierenden sei bestenfalls moderat (r = .29).
- Die niedrigste relative Ähnlichkeit lag zwischen den Selbstbeurteilungen und denen der Kollegen vor (r = .15).

In einer Studie von Staufenbiel (2000) korrelierten Selbst- und Studierendenurteile bei vier Skalen (*Planung und Darstellung*, *Umgang mit Studierenden*, *Interessantheit und Relevanz* sowie *Schwere und Umfang*) zwischen r = .19 und r = .35.

Diskussion

Grundsätzlich stellen Abrami, d'Apollonia und Cohen (1990) fest, dass die Ergebnisse studentischer Lehrveranstaltungsevaluation valide sein können, wenn sie die Meinung der Studierenden zur Lehrqualität korrekt reflektieren, ohne dass diese Ergebnisse das studentische Lernen reflektierten. Auf diese Art würden studentische Lehrevaluationsergebnisse eher selten gesehen, dafür aber häufig als Maß *effektiver Lehre*.

Somit ist festzuhalten, dass, wenn studentische Lehrevaluationsinventare die Zufriedenheit Studierender hinsichtlich der Durchführung einer Lehrveranstaltung erheben sollen, Lernerfolg kein angemessenes Kriterium darstellt. Abschlussnoten beziehungsweise Lernerfolg sind somit kein Kriterium im Sinne eines direkten Maßes (im Sinne von Anastasi, 1984) einer studentischen Einschätzung der Lehrqualität.

Auch die Selbst- und Fremdbeurteilungen der Dozenten und Beobachter sind nicht als direktes Maß der Zufriedenheit von Studierenden anzusehen.

Messick (1989a) gibt unabhängig vom Forschungsgegenstand zu bedenken, dass auch Kriterien eine zu evaluierende Messung sind und somit von konstruktirrelevanter Varianz kontaminiert werden können, wie zum Beispiel durch selektive Aufmerksamkeit oder Halo-Effekte.

Als Schlussfolgerung kann somit festgehalten werden, dass mangels eines konkreten Kriteriums diese Validitätsart als nicht angemessen für einen Validitätsnachweis studentischer Lehrevaluationsinventare im Sinne einer Zufriedenheitsmessung angesehen werden kann.

5.6.2 Inhaltsvalidität

Bedeutung

Gegebene Inhaltsvalidität bei studentischen Lehrevaluationsinventaren hängt maßgeblich von folgenden Aspekten ab: Ein klares Verständnis von dem, was *effektive Lehre* ist, sei eine klare Voraussetzung für die Konstruktion von Lehrevaluationsinventaren (Spooren et al., 2013) und die Items eines studentischen Lehrevaluationsinventars sollten laut Marsh (2007) mit den generellen Prinzipien von Lehre und Lernen übereinstimmen. Diese Prinzipien sollten einen Schwerpunkt auf Theorie und Forschung in der Erwachsenenbildung besitzen, da diese besonders im Kontext Höherer Bildung relevant seien.

Beispiel

Beispielsweise wurden bei der Konstruktion des Englischsprachigen Inventars *Students' Evaluations of Educational Quality (SEEQ)* zur Sicherstellung der Inhaltsvalidität folgende Kriterien verwendet (Marsh & Dunkin, 1996): Ein großer Item-Pool wurde aus einer Literaturrecherche und aus geführten Interviews mit Angestellten und Studierenden gewonnen. Dabei sollten sie angeben, was sie als effektive Lehre ansehen. Weiterhin wurden Studierende und Angestellte gebeten, die Relevanz der Items zu bewerten: Angestellte beurteilten, ob die Items nützlich für ein Feedback seien, und anhand offener Kommentare von Studierenden wurde untersucht, ob wichtige Aspekte ausgelassen wurden. Weiterhin zeige ein Vergleich mit denen von Feldman (1976) genannten Kategorien, dass diese sich substanziell mit denen des SEEQ überschneiden. Weiterhin wurden generelle Prinzipien des Lernens von Erwachsenen (zum Beispiel Mackie, 1981) mit den Inhalten des SEEQ erfolgreich verglichen.

Diskussion

Spooren et al. (2013) kritisieren, dass viele Inventare ohne eine klare Theorie effektiver Lehre konstruiert worden seien. Somit wäre Inhaltsvalidität nicht gegeben, und Inventare würden daher nicht dem Anspruch entsprechen, zu messen, was sie messen sollen.

Auch Ory und Ryan (2001) kritisieren, dass viele Institutionen die Zieldomäne effektiver unterrichtlicher Charakteristika oder effektiven unterrichtlichen Verhaltens nicht definiert hätten und somit die ausgewählten Items für Evaluationsinventare mangelhaft seien. Somit seien laut Onwuegbuzie, Daniel und Collins (2009) Item- als auch Sampling-Validität bedroht.

Weiterhin sehen Spooren et al. (2013) und Onwuegbuzie et al. (2009) eine weitere Bedrohung der Inhaltsvalidität: Da die Inventare in der Regel von Angehörigen der universitären Leitung (Administrators) konstruiert würden, könne sich deren Perspektive was effektive Lehre betrifft, nicht immer mit derjenigen der Studierenden decken. Deutlich wird das an der schon im Abschnitt 3.3.2 beschriebenen Studie von Onwuegbuzie et al. (2007): Sie untersuchten an einem College unter anderem die inhaltsbezogene Validität eines Lehrevaluationsbogens. Dabei befragten sie Studierende, welche Charakteristika sie bei effektiven College-Dozenten wahrnehmen. Als Ergebnis stellten sich neun Themen heraus. Im Vergleich mit dem

von der Universität ursprünglich eingesetzten Inventar zeigte sich, dass drei dieser neun von den Studierenden als wichtig angesehenen Themen nicht abgefragt wurden: Studierendenzentriertheit, Expertentum und Enthusiasmus. Aber auch wenn die Wahrnehmungen von Studierenden und Dozenten bezüglich effektiver Lehre positiv korrelierten, zeigten sich Unterschiede: Zum Beispiel empfanden die Studierenden die Vorbereitung des Dozenten auf den Unterricht wichtiger als die Dozenten selbst. An diesen Beispielen kann das Problem eines *Confirmation Bias* deutlich werden: Experten bestätigen die Items eines Lehrevaluationsinventares als angemessen, da diese das repräsentieren, was diese selbst unter Lehrqualität verstehen.

Bei Ergebnissen von Studierendenfragebögen sind wahrscheinlich noch weitere Faktoren zu beachten, die zu Verzerrungen der Interpretationen führen können. Laut Kember, Jenkins und Chi Ng (2004) könnten Studierende jeweils dazu tendieren, Items aus der Perspektive heraus zu beantworten, was sie selber unter guter Lehre verstünden: Studierende, die Frontalunterricht beziehungsweise traditionellen Unterricht (didactic teaching) bevorzugten, lehnten Aspekte der Interaktion ab, und bei anderen Studierenden war es genau umgekehrt. Weiterhin zeigten Kember und Wong (2000) an Interviews von Undergraduate-Studierenden in Hong Kong, dass deren Wahrnehmung von guter Lehrqualität als ein Ergebnis der Interaktion von den Überzeugungen anzusehen ist, was sie selbst unter guter Lehre verstehen, und mit dem, was sie glauben, was die Dozenten unter guter Lehre verstehen: Ersteres lässt sich auf einem Kontinuum von aktivem und passivem Lernen abbilden und Letzteres zwischen vermittelndem (transmissive) und nicht-traditionellem Lehren.

Schlussendlich zeigt sich, dass Inhaltsvalidität ein wichtiger Bestandteil der Konstruktion von Lehrevaluationsinventaren sein sollte: Dabei sollten verschiedene Aspekte beachtet werden, so dass die Konstruktion auf einer angemessenen theoretischen Fundierung basiert. Dazu gehören wie eben aufgeführt die Klärung, was genau gemessen werden soll, der Einbezug entsprechender Theorien und die Schlussfolgerungen empirischer Studien.

5.6.3 Faktorielle Validität

In den Übersichten zu Lehrevaluationsinventaren zeigt sich, dass Faktorenanalysen in Validierungsstudien häufig eingesetzt wurden (Braun, 2007, S. 26-49; Rindermann, 2009, Kap. 6; Sippel, 2014). Ziel sei es, Itemsätze zu

strukturieren und die Zusammenhangsmuster zwischen Items auf grundlegende Faktoren zurückzuführen. Diese Faktoren erklärten dann das Zusammenhangsmuster zwischen den Items. (Rindermann, 2009, S. 79)

Solch ein Zusammenhangsmuster wird in Abbildung 7 für den Faktor *Planung und Darstellung* des Fragebogens zur Evaluation von Vorlesungen (Staufenbiel, 2000) veranschaulicht.

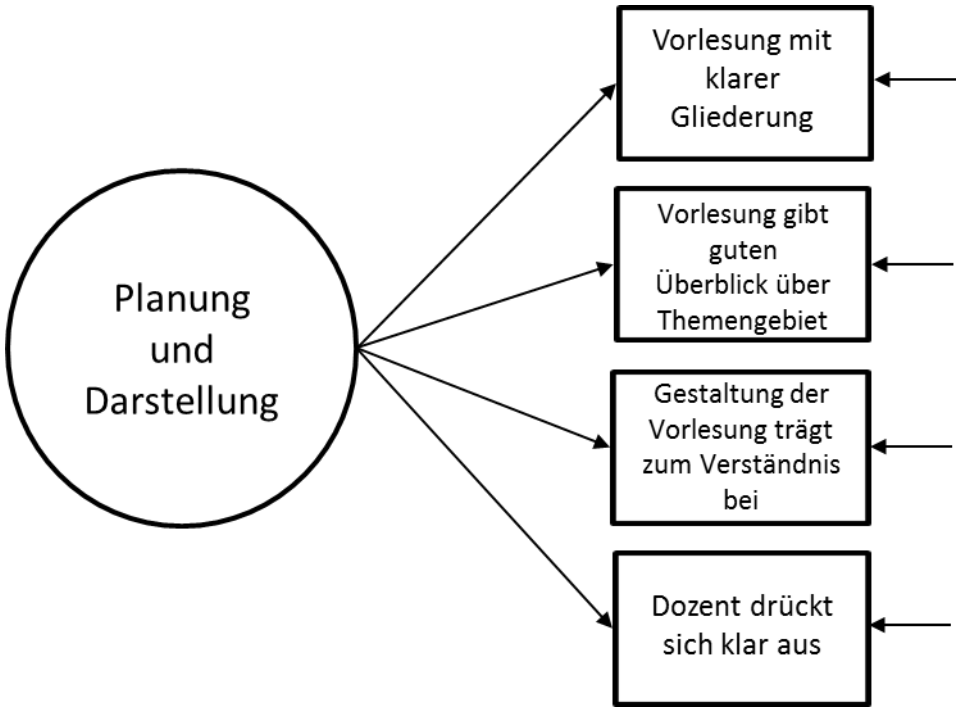

Abbildung 7: Die auf Basis einer Faktorenanalyse identifizierte Skala „Planung und Darstellung“

In Tabelle 1 wird eine Auswahl von Ergebnissen zum Thema der Dimensionalität bei Lehrevaluationsinventaren vorgestellt, um einen Einblick in bisherige Forschungsergebnisse darbieten zu können. Dementsprechend handelt es sich nicht um eine vollständige Übersicht zu allen Inventaren, bei denen Fragen zur faktoriellen Validität untersucht wurden. Beachtet werden sollte, dass sich die Analysen dahingehend unterscheiden, dass sie verschiedene Schätzverfahren (wie Maximum Likelihood, Hauptachsenanalysen oder Hauptkomponentenanalysen), Gütekriterien (unter anderem χ^2-Tests, RMSEA), Abbruchkriterien bei EFAs (Kaiser, Screeplot, Parallelanalysen) angewendet haben und somit nicht immer direkt vergleichbar sind.

Tabelle 1: Überblick über gefundene Dimensionen in Lehrevaluationsinventaren anhand von Faktorenanalysen

Instrument	*Gefundene/replizierte Dimensionen*
Fragebogen zur Veranstaltungsbeurteilung im Studienfach Psychologie (Diehl & Kohr, 1977)	1. Relevanz und Nützlichkeit der Veranstaltungsinhalte: 10 Items 2. Verhalten des Dozenten gegenüber den Veranstaltungsteilnehmern: 10 Items 3. Angemessenheit von Schwierigkeit und Umfang der Veranstaltungsinhalte: 10 Items 4. Methodik und Aufbau der Veranstaltung: 10 Items
Fragebogen der Einstellungen und Verhaltensreaktionen in Seminaren (Müller-Wolf, 1977)	1. Gefühl sozialemotionaler Unterstützung und Zufriedenheit mit dem Seminar 2. kreativ-freies vs. verkrampft-gehemmtes Verhalten 3. Entwicklung von Eigeninitiative
Fragebogen zur Beurteilung von Lehrveranstaltungen (Winteler & Schmolck, 1979; Winteler & Schmolck, 1983)	1. Klima 2. Fachdiskussion 3. Stoffverständnis 4. Wiederholungen 5. Stoffauswahl und Gliederung 6. Schwierigkeit 7. Relevanz-Verdeutlichung Spezifische Faktoren: Akustische Verständlichkeit, Aktivierung, Lehrziele Leerlauf, Beispiele und Rückmeldung
Fragebogen zur Einschätzung der Ausbildungsqualität (Kramis, 1990)	1. Bedeutsamkeit (Inhalte und Ziele) 2. Effizienz (Lernorganisation und Lernaktivitäten) 3. Lernklima
Fragen zur Veranstaltung (Esser, 1994)	1. Bewertung des Dozenten 2. Bewertung der Veranstaltung 3. Verständlichkeit/Überforderung
Marburger Fragebogen zu Akzeptanz der Lehre (Basler, Bolm, Dickescheid & Herda, 1995)	1. Motivation für den Lernstoff 2. Didaktik und Organisation 3. Fachliche Kompetenz 4. Motivation zur aktiven Teilnahme

Fortsetzung auf folgender Seite.

Fortsetzung Tabelle 1:

Lehrverhaltensinventar (Astleitner & Krumm, 1996)	1. Sprache: drei Items 2. Nonverbales Verhalten 3. Erklärung des Lehrstoffes 4. Organisation: vier Items 5. Motivierung: drei Items 6. Aufgabenorientierung 7. Belohnung: drei Items 8. Partizipation: drei Items
Fragebogen zur Beurteilung einer Lehrveranstaltung durch Studierende (Westermann et al., 1998)	1. Anregung zur inhaltlichen Auseinandersetzung 2. Arbeitsaufwand und -belastung der Studierenden 3. Interaktion und soziales Klima 4. Strukturierung der Lehrinhalte 5. Engagement der Lehrenden 6. Prüfungsanforderung und andere extrinsische Motivation 7. Erwerb von Wissen und Verständnis: zwei Items
Fragebogen zur Beurteilung einer Lehrveranstaltung durch Studierende – Weiterentwicklung (Braun, 2002; zitiert nach Braun 2007)	1. Arbeitsklima unter den Studierenden 2. Motivation der Studierenden 3. Soziale Kompetenz der Lehrenden 4. Zielklärung der Lehrveranstaltung 5. Aufwand/Anforderungen: drei Items 6. Lehrveranstaltungsräume/Rahmenbedingungen 7. Didaktische Kompetenz der Lehrenden 8. Engagement der Lehrenden: fünf Items 9. Zufriedenheit mit der Veranstaltung
Fragebögen zur Evaluation von jeweils Vorlesungen, Seminaren oder Praktika (Staufenbiel, 2000)	1. Planung und Darstellung 2. Umgang mit Studierenden 3. Interessantheit und Relevanz 4. Schwierigkeit und Umfang 5.a Qualität der Referate bei Seminaren 5.b Betreuung bei Praktika

Fortsetzung auf folgender Seite.

Fortsetzung Tabelle 1:

Heidelberger Inventar zur Lehrveranstaltungsevaluation II (HILVE-II) (Rindermann, 2009)	1. Lehrkompetenz 2. Anforderung 3. Referate 4. Beteiligung 5. Fleiß 6. Klima 7. Thema 8. Redundanz
Trierer Inventar zur Lehrevaluation (Gollwitzer & Scholtz, 2003)	1. Anregung und Motivation 2. Strukturiertheit und Didaktik 3. Interaktion und Kommunikation 4. Persönlicher Gewinn durch die Veranstaltung 5. Anwendungsbezug
Fragebogen zum Dozierendenverhalten (Koch, 2004)	19 Faktoren nicht reproduzierbar

Diskussion

Onwuegbuzie et al. (2007, S. 118-119) kritisieren die alleinige Verwendung exploratorischer Faktorenanalysen, da sie dazu führe, dass Items in einem Inventar nicht mehr Charakteristika effektiver Lehre, sondern Dimensionen eines gegebenenfalls nicht auf theoretischer Basis entwickelten Instruments repräsentierten. Ory und Ryan (2001, S. 34-35) schlussfolgern, dass eine faktorenanalytische Vorgehensweise der Analyse hunderter Mathematikaufgaben gleiche, die dann antwortbasiert gruppiert würden, und diese Cluster dann als essenzielle Fähigkeiten zur Lösung mathematischer Probleme angesehen würden. Das bedeute, dass Items in Inventare aufgenommen würden, weil Studierende sie ähnlich beantworteten und nicht, weil sie eine gezielt erhobene Eigenschaft darstellten.

Abrami (1989) kritisierte, dass Faktorenanalysen inkonsistente Ergebnisse aufwiesen und daher nicht unbedingt zu verwenden seien. Dies zeigte sich auch in der Übersichtsarbeit von Braun (2007, S. 26-49), bei der sieben von elf Inventaren uneindeutige Faktorlösungen, nicht replizierte Strukturen oder ähnliches aufwiesen.

Da die Diskussion zur Verwendung von Faktorenanalysen bei studentischen Lehrevaluationsinventaren komplex ist, wird diese in einem breiteren Kontext in Abschnitt 5.7.2 weitergeführt.

5.6.4 Klassische Testtheorie

Bedeutung

Im Rahmen der Klassischen Testtheorie wurde insbesondere die interne Konsistenz der Lehrevaluationsinventare überprüft. Hiermit soll sichergestellt werden, dass die Skalen und Subskalen weitgehend messfehlerfrei die Einschätzungen der Studierenden bezüglich der Lehrqualität widergeben.

Beispiele

Braun (2007, S. 26-49), Rindermann, (2009, Kap. 7) und Sippel (2014) haben interne Konsistenzen verschiedener deutschsprachiger Lehrevaluationsinventare zusammengetragen (eine Auswahl von diesen wird in Tabelle 2 wiedergegeben):

Tabelle 2: Interne Konsistenz einer Auswahl von Lehrevaluationsinventaren

Inventar	*Gesamtskala*	*Subskalen*
Fragebogen zur Veranstaltungsbeurteilung im Studienfach Psychologie (Diehl & Kohr, 1977) und Nachuntersuchung von (Kleine & Merkens, 1979)	-	.91 - .95
Fragebogen zur Beurteilung von Lehrveranstaltungen (Winteler & Schmolck, 1979)	Keine Skalenbildung vorgesehen, trotzdem Faktorenanalyse durchgeführt	
Marburger Fragebogen zur Akzeptanz der Lehre (Basler et al., 1995)	-	.69 - .86
Lehrverhaltensinventar (Astleitner & Krumm, 1996)	.24 - .84	
Fragebogen zur Beurteilung einer Lehrveranstaltung durch Studierende (Westermann et al., 1998)	-	-
Fragebogen zur Beurteilung einer Lehrveranstaltung durch Studierende – Weiterentwicklung (Braun, 2002; zitiert nach Braun 2007)	-	.66 - .88
Fragebogen zur Evaluation von Vorlesungen, Seminaren oder Praktika (Staufenbiel, 2000)	-	.50 - .81

Fortsetzung auf folgender Seite.

Fortsetzung Tabelle 2:

Heidelberger Inventar zur Lehrveranstaltungsevaluation II (HILVE-II) (Rindermann, 2009)	-	.01 - .91
Trierer Inventar zur Lehrevaluation (Gollwitzer & Scholtz, 2003)	-	.73 - .84
Fragebogen zum Dozierendenverhalten (Koch, 2004)	-	.60 - .98

Diskussion

Aufgrund verschiedener theoretischer Überlegungen kann in Zweifel gezogen werden, dass interne Konsistenzen eine Aussagekraft bei der Mehrheit von Lehrevaluationsinventaren besitzen. Es besteht sogar die Gefahr, dass im Kontext der Optimierung der Reliabilität, Inventaren unnötig Items zugefügt oder entfernt werden. Da dies ein Gegenstand komplexerer Diskussion ist, wird auch hier im Abschnitt 5.7.2 ausführlich darauf eingegangen.

5.6.5 Generalisierbarkeitstheorie

Bedeutung

Im Kontext studentischer Lehrevaluationsinventare wurde anhand der Generalisierbarkeitstheorie insbesondere folgende Fragestellung untersucht:

Auf welche Varianzquellen lässt sich die Variabilität studentischer Urteile zurückführen, und wie zuverlässig ist das studentische Urteil innerhalb und zwischen verschiedenen Zeitpunkten oder Situationen? Dadurch kann überprüft werden, ob standardisierte Lehrevaluationsinventare unter allen Bedingungen gleich reliabel, und deren Ergebnisse über verschiedene Bedingungen universitärer Lehre hinweg generalisierbar sind.

Beispiel

In einer Studie von Gillmore (1978) wurde die Zuverlässigkeit studentischer Urteile an der Universität von Washington mit dem *Instructional Assessment System* untersucht.

In zwei G-Studien wurden einmal Dozenten und einmal Kurse als Facetten der Diskriminierung spezifiziert. Die studentischen Urteile sollten bei der ersten G-Studie über die Facette der Kurse (geschachtelt in Dozenten) generalisieren und bei der zweiten über Dozenten (geschachtelt in Kurse). Bei beiden wurden als weitere

Facetten der Generalisierung die Items und Studierenden spezifiziert. Eine Varianzkomponente für die Interaktion konnte nicht geschätzt werden.

1. G-Studie: Die Varianzkomponente der Dozenten lag mit .08 im Vergleich zu den anderen in mittlerer Höhe. Die der Kurse (geschachtelt in Dozenten) lag bei .12, die der Items bei .02 und am höchsten die der Studierenden (geschachtelt in Kurse und Dozenten) mit .49 und dem Residuum samt angenommener Interaktionen .50. Weitere Komponenten (zum Beispiel Items geschachtelt in Dozenten) werden der Übersicht wegen hier nicht aufgeführt; sie zeigten auch kein bedeutsames Ausmaß.

2. G-Studie: Die Komponente der Kurse ist sehr niedrig (.01), die der Dozenten (geschachtelt in Kursen) höher (.16). Die Komponente der Items ist wiederum niedrig (.01) und die der Studierenden (geschachtelt in Dozenten und in Kursen) mit .48 sowie das Residuum inklusive Interaktion ist höher (.47). Auch hier werden weitere Komponenten der Übersicht wegen nicht aufgeführt.

In den jeweiligen D-Studien wurden jeweils zwei Generalisierbarkeitskoeffizienten geschätzt: Einen zur Generalisierung über alle möglichen Kurse, die ein Dozent halten könnte, über alle Studierenden, die unterrichtet und alle möglichen Items, die zu dem Thema abgefragt werden könnten. Sowie ein weiterer Koeffizient, bei dem neben den Studierenden und Items nur die konkret in die D-Studie eingeschlossenen Kurse (1. Studie) oder Dozenten (2. Studie) berücksichtigt wurden.

1. D-Studie: Um einen nach Ansicht der Autoren der Studie ausreichend hohen Generalisierbarkeitskoeffizienten bei der Generalisierung über alle möglichen Kurse zu erlangen, sollten mindestens fünf Kurse berücksichtigt werden, aber die Menge der Studierenden sei nicht so bedeutsam. Falls nicht über alle möglichen Kurse generalisiert werden sollte, reicht ein Kurs aus.

2. D-Studie: Generalisierbarkeitskoeffizienten würden bei keiner Dozenten-Stichprobengröße eine ausreichende Höhe erlangen, wenn über alle möglichen Dozenten generalisiert werden soll. Wenn nur die in die Analysen

eingeschlossenen Dozenten berücksichtigt werden, sei eine angemessene Höhe unabhängig von den anderen Facetten erreichbar.

Zusammengefasst schlussfolgern die Autoren, dass in der ersten Studie 40% der Varianz auf den Dozenten zurückzuführen sei (Dozent allein, sowie Kurse in Dozent geschachtelt). Bei der zweiten Studie seien nur 6% auf den Kurs zurückzuführen. Weiterhin könne ein Kurs nicht zuverlässig anhand studentischer Urteile evaluiert werden, wenn über alle möglichen Dozenten generalisiert werden solle. Weiterhin gehen sie von einer Interaktion von Kursen mit Dozenten aus. Daher empfehlen die Autoren zur Erhöhung der durch die Studierenden wahrgenommenen Qualität, bestimmte Dozenten bestimmten Kursen zuzuordnen.

Diskussion

Ähnlich wie bei der Klassischen Testtheorie unterliegt die Generalisierbarkeitstheorie bestimmten Annahmen. Ob diese auf die Messung von Lehrqualität durch Studierende zutreffen, wird ebenfalls in Abschnitt 5.7.2 diskutiert.

5.6.6 Item Response-Theorie

Bedeutung

Wie bereits erwähnt, schätzt die Item Response-Theorie Ausprägungen latenter Variablen und bestimmt auch Item-Charakteristika wie die Schwierigkeit. Auch kann durch die Modell-Spezifikation eine Dimensionsprüfung durchgeführt werden: Zum Beispiel, ob alle Items eines Lehrevaluationsinventars Ausprägungen eines gemeinsamen latenten Konstruktes sind, oder die Items jeweils eine dozenten- oder veranstaltungsspezifische Dimension repräsentieren.

Beispiel

In einer brasilianischen Studie (Junior, Fernando de Jesus Moreira, Zanella, Lopes & Seidel, 2015) wurde ein ordinales IRT-Model berechnet: Zunächst wurde eine Faktorenanalyse durchgeführt, nach der 23 verschiedene fünfstufige Lehrevaluations-Items (wie zum Beispiel „Der Dozent regiert auf Anfragen der Studierenden“, „Die Qualität der Lehrmittel ist gegeben“ und „Führt zu Verbesserung der Leistung und zu positiven Ergebnissen“) als eindimensional angesehen und in eine IRT-Analyse überführt wurden. Das latente Konstrukt wurde als *Zufriedenheit der Studierenden mit einem Kurs* definiert und Personenparameter sowie die Itemschwierigkeiten geschätzt. Auf Basis der IRT-Analyse wurden zwei Items entfernt, da ihre Antwortkategorien

nicht ausreichend differenzierten (unter anderem „Der Professor war anwesend und pünktlich“). Die Itemcharakteristische Funktion des Items „Fähigkeit zur Stimulation/ Motivation des Dozenten“ wird in Abbildung 8 dargestellt.

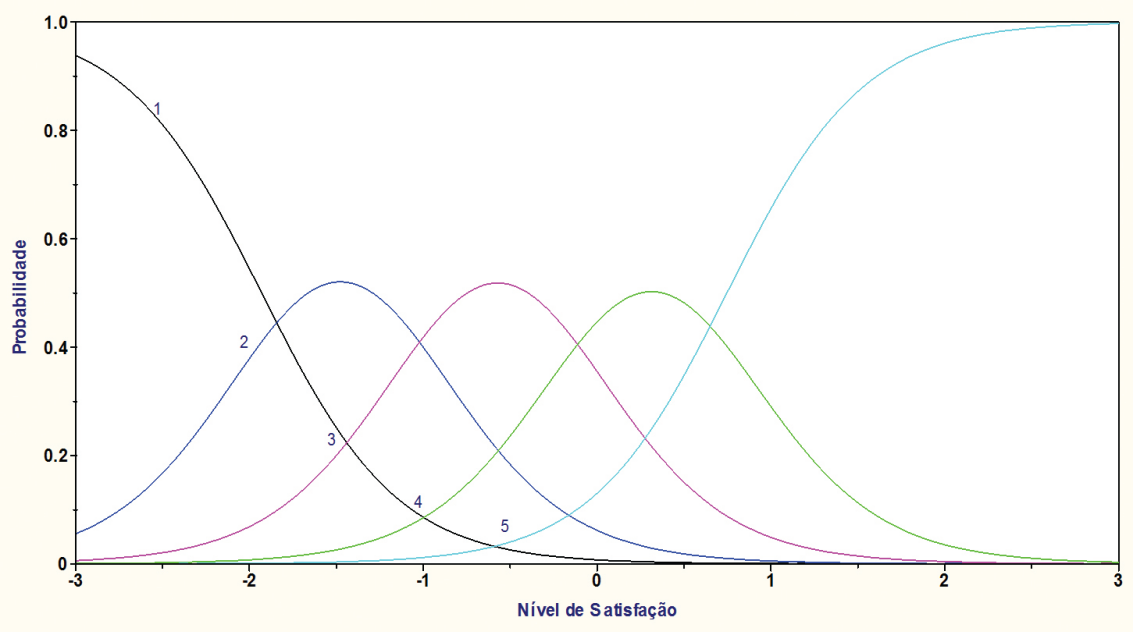

Abbildung 8: Itemcharakteristische Funktion des Items „Fähigkeit zur Stimulation/ Motivation des Dozenten“ (Capacidade de estímulo/motivação apresentada pelo professor). Auf der y-Achse ist die Wahrscheinlichkeit für die jeweilige Antwortmöglichkeit angegeben und auf der x-Achse das Ausmaß der Zufriedenheit der Studierenden; entnommen aus Junior, Fernando de Jesus Moreira et al. (2015, S. 146).

Diskussion

So wie die Klassische Testtheorie, Generalisierbarkeitsteorie und die faktorielle Validität wird die kritische Reflektion zur IRT bezüglich studentischer Lehrevaluationsinventare in Abschnitt 5.7.2 dargestellt.

5.7 Besonderheiten bei der Struktur und Analyse von Lehrevaluationsdaten

Bezüglich der Datenanalyse in Validierungsstudien ergeben sich unterschiedliche Anforderungen, je nachdem was für ein Konstrukt erfasst werden soll, und auch unter welchen Bedingungen dieses gemessen wird. Bei studentischen Lehrevaluationsdaten unterscheiden sich häufig sowohl die Struktur des Konstrukts als auch die Bedingungen, unter denen es gemessen wird von in herkömmlichen Tests gemessenen Merkmalen. Diese Aspekte werden in den folgenden Abschnitten erläutert und diskutiert.

5.7.1 Die hierarchische Struktur studentischer Lehrevaluationsdaten

Allgemein

In der pädagogischen Forschung werden häufig Konstrukte erforscht, die die Dozierenden, die Unterrichtseinheit (wie Schulklassen) oder das Thema betreffen (zum Beispiel dessen Relevanz). Allerdings werden die Daten, die diese Konstrukte messen sollen, häufig nicht direkt, sondern über jeweils mehrere Schüler oder Studierende gemessen.

Wenn nun in Validierungsstudien dieser Umstand nicht beachtet wird und typische Analyseverfahren wie Faktorenanalysen verwendet werden, ergeben sich mathematische Probleme: Faktorenanalysen setzen wie viele andere statistische Verfahren die Unabhängigkeit der einzelnen Messungen (der einzelnen Personen) voneinander voraus. Eine Verletzung dieser Annahme - in Form jeglicher Gemeinsamkeit innerhalb von Gruppen - führe zu unkorrekten Schätzungen von Modellparametern, Standardfehlern und Gütekriterien. (Heck, 2001; zitiert nach Toland, 2005, S. 273)

Eine entsprechende Anwendungsempfehlung lautet, nur Mittelwerte der Evaluationsdaten (wie von Schulklassen) in Faktorenanalysen einzubeziehen (Marsh, 1983; zitiert nach Toland, 2005, S. 273). Allerdings missachte dieser Ansatz die individuelle Variabilität innerhalb einer Klasse. Dadurch erscheinen Beziehungen zwischen Variablen stärker, als sie in Wirklichkeit sind (Kaplan & Elliott, 1997).

Um diesen Problemen zu begegnen, sind *Multilevel Faktorenanalysen*[3] eine angemessene Verfahrensgruppe: Hierbei können individuelle Itemantworten unter Berücksichtigung ihrer Gruppenzugehörigkeit analysiert werden (zum Beispiel Muthén & Bengt O., 1991). Dadurch kann anhand des manifesten Levels der Datenerhebung (wie Schüler oder Patienten) eine Aussage über das latente Level des eigentlich interessierenden Konstrukts (entsprechend Lehrer oder Ärzte) getroffen werden. Dabei werden die latenten Quellen der Varianz auf Gruppen- und Individualebene dekomponiert.

Bezug zu studentischen Lehrevaluationsdaten

Das beschriebene Problem ist im Kontext studentischer Lehrevaluationen relevant: Das interessierende Konstrukt liegt auf der latenten Ebene (beispielsweise „Eignung des Dozenten"), die Datenerhebung beziehungsweise die Indikatoren dieses Konstrukts werden aber auf der Ebene der Studierenden erhoben.

Sengewald und Vetterlein (2015) zeigten anhand des PELVE-Lehrevaluationsinventars (*Prozess- und ergebnisorientierte Lehrveranstaltungsevaluation*) die Überlegenheit einer Multilevel-Faktorenanalyse gegenüber einer herkömmlichen Faktorenanalyse: Im Kontrast zu konventionellen Analysen auf individueller Ebene (keine Gruppierung der Studierenden innerhalb der Veranstaltungsebene) oder der Veranstaltungsebene (jeweilige Aggregation der einzelnen Studentenurteile) zeigte nur die Multilevel-Analyse eine adäquate Modellanpassung.

Dementsprechend sind insbesondere auch bei Lehrevaluationsdaten die Ebenen der Datenerhebung und des eigentlich interessierenden Merkmals zu beachten und ein entsprechendes Analyseverfahren auszuwählen.

Inwiefern Faktorenanalysen grundsätzlich in Bezug auf Instrumente der Lehrevaluations sinnvoll sind, wird im folgenden Abschnitt (5.7.2) diskutiert.

5.7.2 Die Struktur studentischer Lehrevaluationsitems und deren Konstrukt

In der Regel werden Items als Indikatoren des zu messenden Merkmals beziehungsweise Konstrukts betrachtet: Sind die Ausprägungen einer Person auf den

[3]Multilevel-Faktorenanalysen sind nicht mit Faktorenanalysen höherer Ordnung zu verwechseln. In diesen bezieht sich der Begriff „hierarchisch" auf die Korrelation latenter Variablen, die sich durch eine Variable höherer Ordnung erklären lässt: zum Beispiel Mulaik und Quartetti (1997).

Items eines Intelligenztests hoch ausgeprägt, wird von einer entsprechend hohen Intelligenz dieser Person ausgegangen. Auch wenn dies die weit verbreitete Vorstellung von dem Beziehungsmuster zwischen Items und Konstrukt widerspiegelt, sind je nach Testintention und Art des zu messenden Konstrukts noch andere Möglichkeiten in Betracht zu ziehen. Insbesondere Validierungsstudien von Tests und Fragebögen sollten unter diesen Gesichtspunkten kritisch hinterfragt werden.

Reflektive und formative Messmodelle

Das eben genannte Beispiel zur Intelligenzmessung kann exemplarisch für ein *reflektives Messmodell* herangeführt werden: Der Grad der Merkmalsausprägung einer Person reflektiert sich in der Art der Itembeantwortung oder dem Lösen von Testaufgaben.

Eine weitere Annahme über das Beziehungsmuster zwischen Testitems und dem zu messenden Konstrukt ist das *formative Messmodell*: Die entsprechenden Items *formen* die Ausprägungen des zu messenden Konstrukts, zeigen aber nicht die jeweilige Merkmalsausprägung an. Hier kann als Beispiel die *gesundheitsbezogene Lebensqualität* angeführt werden: Einzelne Items in einem entsprechenden Fragebogen können unabhängig voneinander hoch, mittel oder gering ausgeprägt sein ohne die Annahmen bezüglich der Beziehung zu dem Konstrukt zu verletzen. Denn typische Merkmale wie das Vorhandensein von Schmerzen oder Übelkeit können je nach Fall unabhängig vorhanden oder abwesend sein, formen aber gemeinsam die Ausprägungen des Konstrukts der gesundheitsbezogenen Lebensqualität. Die Abwesenheit beider Merkmale kann eine hohe Merkmalsausprägung, das Vorhandensein von jeweils einem eine geringere, und die von beidem eine niedrige bedeuten.

Der Anspruch an die Testitems besteht beim formativen Messmodell somit nicht in einer gemeinsamen Variation und kann nicht durch Faktorenanalysen überprüft werden. Ebenfalls kann der Messfehleranteil nicht bestimmt werden, da die Items nicht die Ausprägungen des gemeinsamen Merkmals messen.

Formative Items können in drei Kategorien aufgeteilt werden: *Kausale Indikatoren, Kovariaten* und C*omposite-Indikatoren*: Bei ersterem formen die Itemantworten die Ausprägungen eines Konstrukts wie eben am Beispiel der Lebensqualität beschrieben, das wiederum durch reflektive Items gemessen werden

kann („Mir geht es gut“). Bei der zweiten Kategorie beeinflussen Kovariaten die Ausprägungen des Konstrukts, wie das Geschlecht oder das Alter. Bei der dritten Kategorie ist kein Anspruch gegeben, das theoretische Konstrukt reflektiv zu messen, sondern die jeweils ausgewählten Testitems sind je nach Theorie oder Testabsicht das Konstrukt selbst. Ein Beispiel hierfür sind Berechnungen des sozioökonomischen Status. (Bollen & Bauldry, 2011)

Die Nichtbeachtung einer formativen Struktur kann schwerwiegende Folgen mit sich bringen: Elemente der Klassischen Testtheorie (interne Konsistenzen und Item-Trennschärfen), die Item Response-Theorie sowie exploratorische und konfirmatorische Faktorenanalysen gehen von der Annahme eines reflektiven Messemodells aus. So kann es dazu kommen, dass aufgrund niedriger Item-Trennschärfen oder niedriger Ladungen in Faktorenanalysen Items fälschlicherweise entfernt werden, und dadurch der Test nicht mehr alle notwendigen Bereiche bei einer Messung abdeckt. Weiterhin können – aus formativer Perspektive unbedenkliche – sich unterscheidende Ergebnisse von Faktorenanalysen Verwirrung und unnötige Debatten über die Theorie von Konstrukten auslösen (zum Beispiel Kieffer, Verrips & Hoogstraten, 2009 zur *oralen gesundheitsbezogenen Lebensqualität* oder bezogen auf studentische Lehrevaluationen Abrami, 1989). Aufgrund der potenziell geringeren Korrelation formativer Items können niedrige interne Konsistenzen bei Analysen angezeigt werden. Solche Ergebnisse können dazu verleiten, den Test nach der Spearman-Brown-Korrektur zu verlängern und somit negative Effekte längerer Tests einzugehen.

Zusammengefasst unterscheiden sich formative und reflektive Items hinsichtlich folgender Aspekte:

1. *Der Kausalität zwischen den Items und dem zu messenden Konstrukt:* Bei reflektiven Items wird deren Ausprägung durch die des Konstrukts bestimmt, und die Ausprägungen formativer Items bestimmen die des Konstrukts.
2. *Der Korrelation zwischen den Items*: Reflektive Items korrelieren hoch miteinander, da sie die Ausprägung eines gemeinsamen Merkmals messen, während keine Bedingung an die Korrelation formativer Items gestellt wird.

Formative und reflektive Items müssen nicht getrennt erfasst werden, sondern können sich gemeinsam auf ein Konstrukt beziehen und dementsprechend in einem

Fragebogen gemeinsam aufgeführt werden: Bei oraler gesundheitsbezogener Lebensqualität sind „Zahnschmerzen“ als formativ und „sich angespannt fühlen“ als reflektiv anzusehen (Kieffer et al., 2009).

Gründe für die „Dominanz des reflektiven Modells“

In zwei Reviews zeigte sich, dass einmal 95 von 102 (Petter, Straub & Rai, 2007 zu Instrumenten zur Messung von *Informationssystemen in der Betriebswirtschaft*) und zum anderen 80% (Eggert & Fassott, 2003 in Artikeln einer Zeitschrift zu *Marketing*) der in Validierungsstudien genutzten Konstrukte fälschlicherweise reflektiv spezifiziert wurden, und die formative Struktur nicht berücksichtigt wurde. Demensprechend geht Eberl (2004; S. 23) von einer „Dominanz des reflektiven Modells“ aus und benennt fälschlich reflektiv durchgeführte Analysen oder Annahmen als *Fehler des Typs F*.

Als Ursache für die Dominanz des reflektiven Modells können historische Gründe in Betracht gezogen werden: Testtheoretische Überlegungen begannen mit Messungen von Persönlichkeitseigenschaften wie Intelligenz, die einer reflektiven Struktur entsprechen. Im weiteren zeitlichen Verlauf wurden die Analyseverfahren auf andere Bereiche außerhalb der Persönlichkeitspsychologie übertragen, deren zu messende Struktur aber nicht mehr einer reflektiven glichen.

Faktorenanalysen, interne Konsistenzen und Item Response-Modelle in der studentischen Lehrevaluation

Wie in Abschnitt 5.6 gezeigt, werden häufig interne Konsistenzen und Faktorenanalysen bei Lehrevaluationsinventaren berechnet und angewandt.

Allerdings unterliegt den Items von Lehrevaluationsinventaren größtenteils keine Annahme der Itemhomogentität: Items, die einmal die Motivation des Dozenten, die Organisation der Veranstaltung und die Relevanz der Inhalte abfragen, können, aber müssen nicht miteinander korrelieren. Auf Basis dieser unterschiedlichen Inhalte ist weder eine Messfehlerbestimmung noch eine Suche nach einer Faktorenstruktur über diese Items hinweg sinnvoll.

Es besteht kein zu überprüfender Anspruch, dass abgefragte Inhalte wie „Die Veranstaltung ist gut organisiert“ und „Ich lerne viel in der Veranstaltung“ Reflektionen eines gemeinsamen Konstrukts sind und ein Summenscore gebildet werden kann (beispielhaft entnommene Items aus dem HILVE 2: Electric Paper -

Gesellschaft für Softwarelösungen, 2004). Daher kann eine Person mit hohem Vorwissen wenig lernen, die Veranstaltung aber trotzdem als gut organisiert ansehen. Dementsprechend sind Faktorenanalysen konfirmatorischer als auch exploratorischer Art nicht notwendig, um die theoretischen Grundlagen des Inventars zu überprüfen. Dies gilt auch für die in diesem Kontext selten eingesetzte Item Response-Theorie. Auch bei Items mit verwandt wirkendem Inhalt, wie „Die Dozentin/der Dozent spricht verständlich und anregend“ und „Die Dozentin/der Dozent fasst regelmäßig den Stoff zusammen“ besteht kein theoretischer Anspruch auf eine hohe Korrelation beziehungsweise eine Annahme, sie seien Indikatoren eines gemeinsamen Konstrukts.

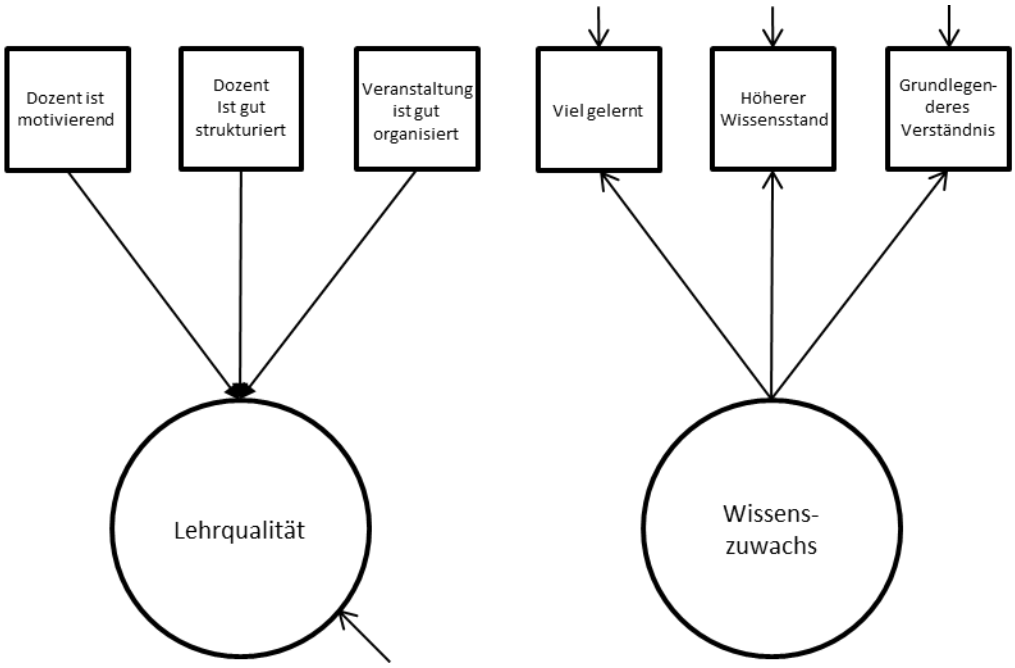

Abbildung 9: Die Konstrukte Lehrqualität mit formativen und Wissenszuwachs mit reflektiven Items

Dagegen kann bei den Items „Ich lerne viel in der Veranstaltung“ und „Mein Wissensstand ist nach der Veranstaltung höher als vorher“ von einem gemeinsamen Konstrukt ausgegangen werden (in etwa „Wissenszuwachs“, siehe Abbildung 9). In diesem Fall besteht die Frage, ob eine Faktorenanalyse - im Sinne einer Überprüfung, inwiefern diese Items Ausprägungen eines gemeinsamen Konstrukts sind – in jedem Fall sinnvoll wäre. Je nach Ziel können Lehrevaluationsdaten auf Einzelitemebene ausgewertet werden, um spezifisch verbesserungswürdige Aspekte zu identifizieren. Die Bildung von Scores ist daher nicht immer notwendig. Als Alternative bei einem Inventar mit vielen Items könnte eine Hauptkomponentenanalyse durchgeführt

werden, um hoch miteinander korrelierende Items zu identifizieren und gegebenenfalls aus Gründen der Sparsamkeit einzelne zu entfernen. Eine Hauptkomponentenanalyse hat den Zweck einer „möglichst umfassenden *Reproduktion* der Datenstruktur durch möglichst wenige Faktoren“ und nicht die Erklärung der Varianz der Variablen. Eine Hauptkomponentenanalyse wird daher häufig nicht als Faktorenanalyse angesehen. (Backhaus, 2008, S. 350, Hervorhebungen im Original)

Anhand des Beispiels reflektiver Items bezüglich eines Konstrukts wie „Wissenszuwachs“ ist zu diskutieren, inwiefern eine Messfehlerbestimmung anhand mehrerer Items überhaupt notwendig ist: Wie schon beschrieben, wurden Faktorenanalysen und Messfehleranalysen häufig in der Persönlichkeitspsychologie eingesetzt. Dabei wurde untersucht, inwiefern verschiedene Aspekte der Persönlichkeit (wie beispielsweise der Geselligkeit oder der Neigung zu Nervosität) die Ausprägungen verschiedener voneinander unabhängiger Konstrukte sind (siehe Big Five in Kapitel 5.6.3 und Abbildung 10). Somit wurden zwei Fragen beantwortet: Ob diese Aspekte Ausprägungen eines gemeinsamen Konstruktes sind, und ob bei einer Messung dieses Konstrukt ausreichend messfehlerfrei gemessen wurde. Bei dem Konstrukt „Wissenszuwachs“, erhoben anhand studentischer Lehrevaluationsinventare, kann davon ausgegangen werden, dass ein Item wie „Ich lerne viel in der Veranstaltung“ ausreicht, um die gewünschten Informationen zu erhalten.

Weiterhin besteht die Gefahr, dass Testnutzer in ihrem Antwortverhalten durch redundante Items beeinflusst werden: Die Wiederholung sehr ähnlicher Fragen kann in unterschiedlichen Interpretationen des Inhalts münden und dadurch in unterschiedliche Antworten. Der Nutzer könnte glauben, dass eine zweite Frage mit ähnlichem Inhalt eine Aufforderung beinhalte, neue beziehungsweise andere Information anzugeben. (Schwarz, 1996)

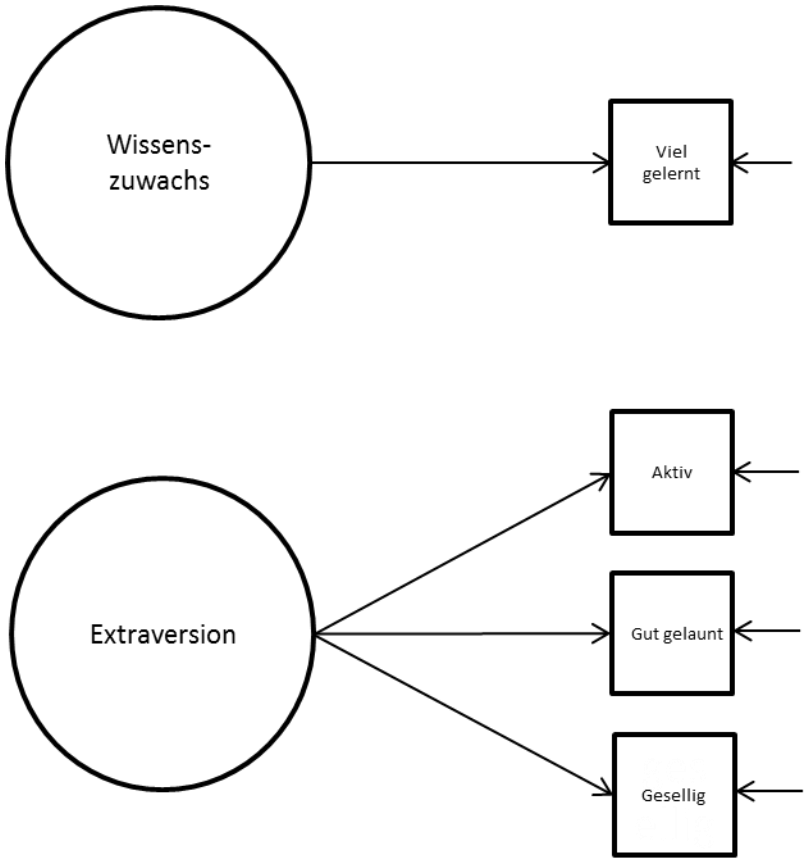

Abbildung 10: Indikatoren verschiedener Konstrukte

Generalisierbarkeitstheorie

Anhand der Generalisierbarkeitstheorie (GT) soll überprüft werden, ob ein als über verschiedene Bedingungen konstant angenommenes Merkmal durch einen Test weitgehend unbeeinflusst von verschiedenen systematischen und unsystematischen Messfehlerquellen erfasst wird (siehe Kapitel 5.6.5). Testwert-Unterschiede eines Individuums sind nach der GT in verschiedenen Messsituationen von einer oder mehreren Fehlerquellen abhängig und unterliegen nicht systematischen Veränderungen aufgrund von Reife oder Lernen. (Shavelson & Webb, 1991)

Doch diese beschriebenen Grundvoraussetzungen der GT treffen auf viele Situationen der Lehrevaluation nicht zu:

1. *Reife oder Lernen:* Auf Basis der genannten Vorannahmen müsste auch die studentische Beurteilung eines Dozenten – wenn der Dozent die Facette der Diskriminierung darstellt - theoretisch stabil über verschiedene Zeitpunkte hinweg bleiben. Allerdings können sich diese aufgrund von gesammelter

Lehrerfahrung und des auf Basis von Rückmeldungen erwünschten Feedbacks verbessern. Auch könnte man annehmen, dass Studierende durch Erfahrung die Qualität von Lehre besser einschätzen können, oder sich ihr Bedarf ändert. Somit sind sie als Facette der Diskriminierung ebenfalls nicht geeignet.

2. *Testwert-Unterschiede nur von Fehlerquellen abhängig:* Auch müsste eine studentische Beurteilung eines Dozenten nur von dessen Kompetenz abhängen, und alle weiteren Einflüsse seien als systematische oder unsystematische Messfehler anzusehen. Allerdings wäre es plausibel, dass einzelne, grundsätzlich schwerer zu vermittelnde Themen zu einem höheren Ausmaß einer entsprechenden Facette der Generalisierung führen, oder dass Studierende auf einen bestimmten Lehrstil unterschiedlich ansprechen. Auch die Studierenden wären wiederum als Facette der Diskriminierung ungeeignet, denn deren individuelle Urteile können sich je nach Dozent und Thema sowie deren Interaktion voneinander unterscheiden. Diese stellen keine Fehlerquellen dar, sondern sind plausible Einflüsse auf das studentische Beurteilungsergebnis.

Zusammengefasst sprechen höhere Einflüsse anderer Varianzquellen nicht gegen das Messinstrument, und der Generalisierbarkeitskoeffizient würde keine Aussage über die Güte eines Tests zulassen. Daher ist keine Aussage über die Zuverlässigkeit eines einzelnen Ergebnisses möglich, da eine Variation über verschiedene Bedingungen plausibel ist. Als Alternative stellt sich die Anwendung einer Varianzkomponenten-schätzung ohne die Vorannahmen der GT dar: Bei jedem Lehrevaluationsitem können spezifischen Annahmen über die Plausibilität der Ausprägung der einzelnen Komponenten aufgestellt und überprüft werden.

Schlussfolgerungen

Itemhomogenität ist eine grundlegende Eigenschaft von Tests, die anhand von Faktorenanalysen überprüft wird und zur Ermittlung des Messfehleranteils in Form interner Konsistenzen berechtigt. Dementsprechend ist bei den theoretischen Überlegungen bei der Testkonstruktion zu bedenken, ob die jeweiligen Items tatsächlich Ausprägungen eines ihnen gemeinsamen Konstrukts sind, beziehungsweise ob sie es überhaupt sein sollen. Falls dies nicht der Fall ist, sind Faktorenanalysen und interne Konsistenzen nicht angemessen, und die Testkonstrukteure sollten die empirische Überprüfung anderer Ansprüche an ihren Test angehen.

Wie gezeigt, ist Itemhomogenität in der Regel bei Lehrevaluationsinventaren nicht beabsichtigt. In seltenen Fällen sind Items mit redundanten Iteminhalten gegeben, aber meistens werden heterogene qualitätsrelevante Facetten einer Lehrveranstaltung abgefragt. Somit sind Überlegungen hinsichtlich der ausreichenden Höhe von Cronbach α sowie reproduzierte und inhaltlich sinnvolle Faktorlösungen im Sinne reflektiver Items nicht notwendig. Probleme, wie die von Abrami (1989) erwähnten inkonsistenten Ergebnisse von Faktorenanalysen, und die bei der Übersichtsarbeit von Braun (2007, S. 26-49) gezeigte häufig vorkommende mangelnde Reproduzierbarkeit angedachter Faktorenstrukturen sind eine Folge der unangemessenen Annahme von Itemhomogentität.

Auch sollte grundlegend bei Validierungsprozessen bedacht werden, dass unabhängig davon, ob reflektive oder formative Items vorliegen, bei einer beabsichtigten Präsentation von Ergebnissen auf Einzelitemebene, Faktorenanalysen und interne Konsistenzen nicht notwendig sein müssen.

5.8 Schlussfolgerung zur herkömmlichen Validitätsüberprüfung studentischer Lehrevaluationsergebnisse

Wie an den Beispielen im Abschnitt 5.6 gezeigt, wurden die genannten testtheoretischen Vorgehensweisen zur Überprüfung von Validität und Reliabilität einzeln oder auch in Kombination in verschiedenen Studien angewandt. Wie bei der Kriteriumsvalidität gezeigt, kann es generell schwierig sein, Validitätsnachweise in Form von Kriterien für studentische Lehrevaluationsergebnisse zu finden. Ein in Hinsicht der Inhaltvalidität angemessen konstruiertes Inventar ist weiterhin bedeutsam, aber viele Inventare scheinen nicht unter ausreichender Berücksichtigung von Aspekten guter Lehre konstruiert worden zu sein. Auch besteht die Gefahr, dass Aspekte nicht berücksichtigt wurden, die Studierende als wichtig erachten. Weiterhin lassen sich angenommene Faktorstrukturen nicht immer replizieren. Eine Erklärung dafür ist, dass die Struktur vieler Lehrevaluationsinventare keinem reflektiven Messmodel entspricht und somit einheitliche und konsistente Faktorlösungen unwahrscheinlich sind. Die Grundlage des reflektiven Modells liegt auch der KTT, IRT und GT zu Grunde, die somit bei vielen Lehrevaluationsitems als Analyseverfahren nicht angebracht sind.

Somit ist bei Lehrevaluationsinventaren als Validitätsart nur die der Inhaltsvalidität als grundsätzlich angemessen zu betrachten. Allerdings ist davon auszugehen, dass Lehrevaluationsinventare noch weitere Ansprüche erfüllen sollten. Im nächsten Kapitel soll diskutiert werden, ob die bislang hier präsentierte Perspektive auf das, was Validität ausmacht, inhaltlich und praktisch sinnvoll ist, und inwiefern alternative Perspektiven und Ansätze für studentische Lehrevaluationsinventare angemessen sind.

6. Argumentationsbasierte Validitätsansätze

6.1 Veränderung des Validitätsverständnisses

In dem beschriebenen *klassischen Validitätsansatz* stehen mehre (maßgeblich Kriteriums-, Inhalts- und Konstruktvalidität) Validitätsarten getrennt nebeneinander, und Validität wird grundsätzlich als Testeigenschaft angesehen. Der Prozess der Validierung an sich beinhaltet hauptsächlich eine binäre Aussage darüber, ob ein Test nun als valide angesehen werden kann, oder kein Nachweis erbracht werden konnte.

Im Laufe der Entwicklung änderten sich diese Perspektiven beziehungsweise wurden sie von einigen Autoren zur Disposition gestellt: Zum einen, dass es nicht getrennte Arten von Validität gebe, sondern Validität als einheitliches Konzept anzusehen ist. Zum anderen, dass nicht die Validität des Tests an sich, sondern die der Interpretationen und der Verwendungen von Testwerten zu überprüfen ist. Auch wie der Prozess der Validierung inhaltlich ablaufen sollte, wurde zur Diskussion gestellt. Im Folgenden werden alle drei Entwicklungslinien beschrieben.

6.1.1 Validität als einheitliches Konzept

Dunnette und Borman (1979) gaben zu bedenken, dass die Annahme verschiedener Validitätsarten zu Verwirrung und grober Vereinfachung führe: Testentwickler legten dadurch ihren Schwerpunkt zu stark auf die Entscheidung, welche Art von Validität zu untersuchen sei, statt genau zu spezifizieren, warum sie einen Test benutzen möchten, und welche Schlüsse sie aus ihm ziehen wollen.

In ähnlicher Weise wies Guion (1980) auf die Gefahr hin, dass, wenn man eine Art von Validität nicht nachweisen könne, man es dann bei den anderen Arten probieren könnte.

Anastasi (1986) ergänzte, dass die Trennung in verschiedene Arten von Validität dazu führte, dass Testkonstrukteure sie wie eine Checkliste abhakten und in Veröffentlichungen berichteten - unabhängig davon, was der Zweck oder die Art des Tests sei. Über die Validitätsarten hinaus seien ihrer Meinung nach aber nahezu alle Informationen im Entwicklungsprozess und der Anwendung für die Validität relevant.

Messick (1989a) argumentierte, dass weder der Nachweis von Inhalts- noch Kriteriumsvalidität ausreichend sei: Inhaltsvalidität beschäftige sich nicht mit dem Einfluss von Methodenvarianz, Antwortprozessen, internen und externen

Teststrukturen, den Unterschieden im Testverhalten über verschiedene Gruppen und Kontexte hinweg, dem Ansprechen von Testwerten auf experimentelle Interventionen oder mit den sozialen Konsequenzen. Auch Kriteriumsvalidität beziehe sich nur auf Evidenz in Form von Korrelationen zwischen Tests und Kriterien und prüfe die genannten Aspekte nicht.

6.1.2 Validität der Testwertinterpretation und ihrer Verwendungen

Wie eben dargelegt, sprechen inhaltliche wie praktische Aspekte gegen die Verwendung verschiedener Validitätsarten. Der zweite Kritikpunkt lautete, dass nicht ein Test an sich valide sein könne, sondern die Interpretationen und die Verwendung der resultierenden Testwerte.

Als Begründung hierfür führt Kane (2013) an, dass Tests nicht als Selbstzweck existierten, sondern die aus ihrer Anwendung resultierenden Werte für Deutungen beziehungsweise für Entscheidungen genutzt werden sollten. Für Werte ein und desselben Tests können verschiedene Interpretationen oder potenzielle Verwendungen möglich sein und sogar in Konkurrenz zu einander stehen: Ein Beispiel hierfür könnte die Frage sein, ob ein hoher Wert eines Tests für die Diagnose einer Depression oder Dysthymie steht, die jeweils unterschiedliche Entscheidungen hinsichtlich einer Therapie implizieren. Eine Aufgabe sei es, die Interpretation mit der höchsten Plausibilität anhand empirischer Evidenz zu identifizieren.

Somit wird die Interpretation eines Testwerts (zum Beispiel höherer Wert steht für höhere Ausprägung eines Persönlichkeitsmerkmals), beziehungsweise die Verwendung der Werte (zum Beispiel kommt jemand ab einem bestimmten Wert für eine berufliche Anstellung in Frage), die sich aus einem Test oder Fragebogen ergeben, auf ihre Validität hin überprüft.

6.1.3 Validierung als Argumentation

Zusätzlich zu den beiden genannten Entwicklungen wurde auch die Veränderung der Perspektive diskutiert, was unter Validierung verstanden werden sollte: Zum Beispiel schlug Cronbach (1988) eine Validitäts-Argumentation (Validity Argument) statt Validitäts-Forschung (Validity Research) vor. Diese Argumentation solle im Sinne einer Evaluation verstanden werden und Konzepte, Evidenz, soziale und personenbezogene Konsequenzen und Werte miteinander verbinden. Eine Validierung

könne niemals abgeschlossen sein, aber trotz unvollständiger Informationen könne eine begründbare Argumentation geführt werden.

Testentwickler sollten sich als Debattierer verstehen: Die Aufgabe der Validierung bestehe nicht darin, einen Test, eine Praxis oder eine Theorie aufrecht zu erhalten. Testentwickler sollten Argumente von Pro und Contra erfassen und für jede Seite die Stärken und Schwächen ihrer Position aufzeigen. Sie sollten klären können, was eine Messung bedeute und die Grenzen jeder Testwert-Interpretation aufzeigen. Diese Interpretationen sollten die Form einer Beschreibung, Vorhersage oder einer empfohlenen Entscheidung annehmen.

Argumentationen müssten Fragen nachgehen, ob ein Test Konsequenzen in der praktischen Anwendung besitzt und seinen Zweck erfüllt. Weiterhin müsste sie Fakten und Unsicherheiten aufzeigen und somit für die Akzeptanz oder Ablehnung einer Praxis Überzeugungsarbeit leisten.

6.1.4 Validitätsdefinition von Messick

Die Entwicklung zu einem einheitlichen Validitätskonzept und zu der Perspektive der Validität von Testwert-Interpretationen und -Verwendungen wurde 1989 in einer Validitäts-Defintion von Messick (1989a, S. 5) zum Ausdruck gebracht: „Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment“ (Hervorhebungen im Original). Somit sollten nach Messick nicht die Beobachtungs- oder Testinstrumente validiert werden, sondern die von den Testwerten abgeleiteten Schlüsse oder andere Indikatoren. Diese Schlüsse bezögen sich auf die Bedeutung oder Interpretation der Testwerte und die Implikationen für Handlungen, die die Interpretationen nach sich ziehen.

Auch Messick erwähnt die Form einer Validitäts-Argumentation, denn die Frage nach der Validität werde durch die Bewertung der Bedeutung oder Balance der Evidenz und gegebenenfalls widersprüchlicher Argumente beantwortet.

6.2 Argumentationsbasierte Validitätsansätze

Die Gruppe der *argumentationsbasierten Validitätsansätze (argument-based approaches to validation)* integrieren die Aspekte eines einheitlichen

Validitätskonzepts sowie der Validierung von Testwert-Interpretationen oder -Verwendungen. Diesen Ansätzen ist gemein, dass sie sich explizit mit der Verbindung zwischen dem Testverhalten eines Probanden und den auf den entsprechenden Testwerten basierenden - je nach Ansatz - Interpretationen, auf diesen basierenden Entscheidungen und Verwendungen sowie deren Konsequenzen beschäftigen, und sie können für quantitative wie qualitative Daten gelten (Bachman, 2015). Die Kernidee sei, die angenommene (proposed) Interpretation und Verwendung von Testwerten explizit und detailliert darzulegen, und dann die Plausibilität dieser Annahmen (Proposals) zu evaluieren (Kane, 2013).

Laut Kane (1992) hat der Begriff des *argumentationsbasierten Ansatzes* den Vorteil, dass er einen *Ansatz* beschreibe und keine *Art von Validität* repräsentiere. Der Begriff der *Argumentation* impliziere ähnlich wie bei Cronbach (1988), dass es eine Zielgruppe gebe, die es zu überzeugen gelte.

In der Gruppe der argumentationsbasierten Validitätsansätze haben sich im Verlauf der letzten Jahrzehnte mehrere Sichtweisen entwickelt, die sich in unterschiedlichem Ausmaß voneinander unterscheiden, Gemeinsamkeiten besitzen und aufeinander aufbauen. Im Folgenden sollen diese Ansätze skizziert und der für diese Arbeit angemessene gewählt werden.

6.2.1 Konstrukt-Modell/Konstruktvalidität

Der historisch älteste Ansatz ist der der *Konstruktvalidität*. Bei ihm handelt es sich im weiteren Sinne um einen argumentationsbasierten Ansatz, da erstmals zu prüfende theoretische Vorhersagen einen festen Bestandteil darstellten und den Ausganspunkt für die später entwickelten Ansätze bildeten.

Der Ansatz der Konstruktvalidität durchlief in seiner Entwicklung zwei Phasen: Zunächst wurde er als eine Art von Validität neben Inhalts- und Kriteriumsvalidität betrachtet (Anastasi, 1986) und im weiteren Verlauf sollte Konstruktvalidität die beiden anderen Validitätsarten und jegliche weitere Evidenz, die sich auf die Interpretation und Bedeutung von Testwerten bezieht, in sich integrieren (Messick, 1989a). Dies äußerte sich auch in den *Standards for Educational and Psychological Testing* von 1999 (AERA, APA & NCME, 1999), in denen Validität allgemein mit Konstruktvalidität gleichgesetzt wurde. Beide Entwicklungsstufen werden im Folgenden erläutert.

Konstruktvalidität als Validitätsart

Cronbach und Meehl (1955) beschrieben, dass Konstruktvalidität zu überprüfen sei, wenn ein Test als Maß eines Merkmals (attribute) oder einer Eigenschaft (quality) interpretiert werde, welches oder welche nicht operational definiert sei. Konstruktvalidität solle untersucht werden, wenn ein Kriterium oder das „universe of content“ (im Sinne von Inhaltsvalidität) als vollkommen unangemessen gelten, die Eigenschaft zu messen.

Das zu lösende Problem eines Wissenschaftlers laute, das Konstrukt zu identifizieren, das die Varianz eines Testverhaltens erkläre. Dieses Konstrukt werde implizit durch seine Rolle in einer Theorie definiert. Falls ein empirischer Beleg für die entsprechende Theorie nicht erbracht werde, sei entweder die Theorie falsch, oder der Test könne das Konstrukt nicht messen. Hierbei spiele faktorielle Validtät als Vorbedingung eine Rolle, denn extrahierte Faktoren gelten als Konstrukt.

Die angenommene Interpretation - in dem Sinne, dass ein Test ein Maß eines bestimmten Merkmals sei - generiere spezifische, testbare Hypothesen. Diese Hypothesen seien ein Mittel, die Behauptung (Claim), dass ein Test ein Konstrukt misst, zu bestätigen oder zu widerlegen. Um solch einen Anspruch zu validieren, müsse ein *nomologisches Netz* existieren. (siehe Abbildung 11): Hierbei werden beobachtbare und theoretische Eigenschaften mit beobachtbaren oder verschiedene theoretische Eigenschaften miteinander in Beziehung gebracht. Somit sei das interessierende Konstrukt in seinem theoretischen Kontext konkret an empirischen Daten testbar.

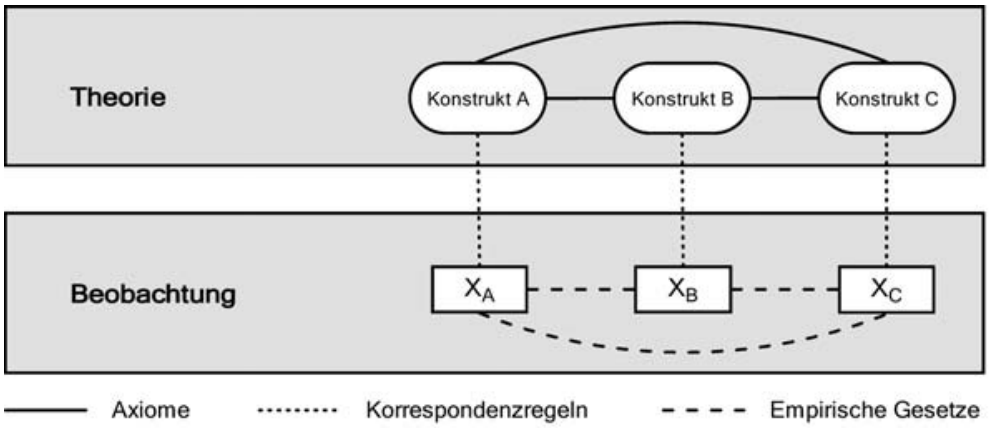

Abbildung 11: Nomologisches Netz aus Hartig und Frey (2007, S. 146)

Das aus dieser Perspektive heraus entwickelte Rahmenmodell für eine Validierung beinhaltet drei Prinzipien (Cronbach; 1971 zitiert nach Kane, 2013, S. 7):

1. Eine angenommene Interpretation muss klar spezifiziert sein.
2. Die Interpretation muss konzeptuell und empirisch evaluiert werden.
3. Es muss berücksichtigt werden, dass alternative Interpretationen mit der aktuell in Betracht gezogenen konkurrieren.

Campbell und Fiske (1959) führten die Begriffe der *konvergenten* und *diskriminanten Validität* ein. Ihre Arbeit handelt maßgeblich von der Frage nach der Eignung eines Tests als Maß eines Konstrukts. Allerdings stellten sie auch beide Begriffe als erforderlich für die Begründung von Konstruktvalidität dar. Konstruktvalidität, beziehungsweise die entsprechenden Annahmen, werden empirisch über Zusammenhangsstrukturen geprüft: Wird von einem möglichst hohen Zusammenhang aufgrund theoretischer Vorannahmen ausgegangen, und dieser auch bestätigt, wird von konvergenter Validität gesprochen, und bei einem niedrigen von diskriminanter Validität.

Konstruktvalidität als integrierendes Validitätskonzept

Wie erwähnt, wurde im weiteren historischen Verlauf Validität als einheitliches Konzept angesehen und mit Konstruktvalidität als integrierender Form gleichgesetzt. Auf Basis dieser Perspektive speist sich nach den Standards von 1999 ein Validitätsnachweis der Testwertinterpretationen aus fünf Quellen von Evidenzen (AERA et al., 1999):

1. Aus dem Inhalt (Content Evidence): unter anderem dass die Items als repräsentativ für die zu erfassende Domäne angesehen werden
2. Aus dem Antwort-Prozess (Response Process): dass alle Fehlerquellen möglichst kontrolliert oder eliminiert werden
3. Aus der internen Struktur (Internal Structure): beinhaltet die statistischen und psychometrischen Charakteristika der Aufgaben, die Skaleneigenschaften und das psychometrische Modell
4. Aus der Beziehung mit anderen Variablen (Relationship to other variables): jegliche auf Korrelationen basierende Evidenzen

5. Aus den Konsequenzen (consequences): die Konsequenzen durch das Bewertungsergebnis und Entscheidungen hinsichtlich der Testpersonen, Institutionen und Gesellschaft

Hinsichtlich der Konstruktvalidität wird zwischen einem *starken* (*strong*) und einem *schwachen* (*weak*) Programm unterschieden (Cronbach, 1988): Das starke Programm bezieht sich auf die ursprüngliche Publikation von Cronbach und Meehl (1955) zur Konstruktvalidität, nach der die theoretischen Ideen so explizit wie möglich dargestellt werden, um diese anhand plausibler konkurrierender Hypothesen bedacht herauszufordern. Das schwache Programm sei reiner erkundender Empirismus, nach dem jegliche Korrelation eines Testwerts mit einer Variable berücksichtigt würde.

Beispiel

Benson (1998) veranschaulichte das starke Programm für den Bereich des Messens von *Prüfungsangst*, für die es starke theoretische Vorannahmen gebe. Benson bezieht sich auf Loevinger (1957) und Nunally (1967), nach denen das starke Programm drei aufeinander aufbauende Komponenten beinhalte: Die substantive, die strukturelle und die externe Komponente.

1. *Substantive Komponente:* Bei dieser wird die theoretische Domäne des Konstrukts spezifiziert und in Form beobachtbarer Variablen definiert. Hierbei wird unter anderem auch auf vorangegangene Forschung zurückgegriffen. Diese Komponente beinhaltet eine theoretische und empirische Ebene in Form des nomologischen Netzes.
2. *Strukturelle Komponente:* Hierbei wird festgelegt, in welchem Ausmaß die beobachteten Variablen zueinander und zu dem Konstrukt in Beziehung stehen. Methoden für deren Überprüfung sind insbesondere Item-Interkorrelationen, Faktorenanalysen, Generalisierbarkeitstheorie und Item Response-Theorie. Sie geben aber keine Auskunft darüber, was gemessen wird.
3. *Externe Komponente:* Hierbei wird begonnen, den Testwerten eine Bedeutung zu geben. Dies geschieht anhand des Festlegens, inwiefern die Maße eines gegebenen Konstrukts in erwarteter Weise mit den Maßen anderer Konstrukte in Beziehung stehen. Dies wird insbesondere in Form von Gruppen-Unterschieden (group-differentation) und Korrelation mit Ergebnissen anderer

Tests überprüft. Als Methoden werden Korrelationen verwendet, wobei Strukturgleichungsmodelle als geeigneter angesehen werden, insbesondere um konkurrierende Hypothesen gegeneinander zu testen.

Übertragen auf das Feld der Prüfungsangst gibt Benson eine Illustration aller drei Komponenten:

1. *Substantiv:* Für das Konstrukt der Prüfungsangst werden drei theoretische Perspektiven mit jeweils entsprechender Operationalisierung dargestellt. Alle drei überlappen sich inhaltlich und teilen Items.
    - Eine Dimension (*Test-Anxiety-Scale)*
    - Zwei Dimensionen: *Sorge* und *Emotionalität* (*Test-Anxiety Inventory)*
    - Vier Dimensionen: *Sorge, körperliche Erregung, Anspannung, Gedankeninhalte ohne Bezug zu dem Test (test irrelevant thinking)* (*Reactions to Test-Scale*)
2. *Strukturell:* Anhand von Faktorenanalysen kann getestet werden, ob andere Konstrukte wie *Versagensangst*, *Selbst-Wirksamkeit* und *Ablenkung* in die Domäne der Prüfungsangst eingefügt werden können.
3. *Extern:* Hinsichtlich der Gruppen-Unterscheidung könne zum Beispiel die Annahme überprüft werden, ob eine Gruppe, die schon Hilfe wegen Prüfungsangst gesucht hatte und eine, die es nicht tat, einen höheren Mittelwert in einem entsprechenden Test hat.

   Zur Überprüfung des Zusammenhangs mit anderen Maßen können die *Sorge*- und die *Emotionalitäts*-Skalen des zweidimensionalen *Test-Anxiety Inventory* mit anderen Variablen in einem Strukturgleichungsmodell in Beziehung gesetzt werden: Somit könne zunächst getestet werden, ob sie einen negativen Einfluss auf die *Leistung* haben. Weiterhin könne geprüft werden, ob sie von ihrem Einfluss auf die *Leistung* her Mediatoren sind, die wiederum vom *akademischen Selbstkonzept* negativ und dem *Misserfolgsvermeidungsmotiv* positiv beeinflusst würden.

Diese drei Komponenten benötigen laut Benson eine starke psychologische Theorie.

Konstruktvalidität studentischer Lehrevaluationsinventare

Auch in dem Kontext studentischer Lehrevaluation wurde Konstruktvalidität überprüft. Hierfür werden eine empirische Studie und eine Übersichtsarbeit zusammengefasst.

Marsh (1987) nimmt *Effektivität der Lehre* (teaching effectiveness) als das zu messende Konstrukt studentischer Lehrevaluationsinventare an. Konstruktvalidität wurde hierbei im Sinne konvergenter (substanzielle Korrelation mit einer Reihe anderer Indikatoren effektiver Lehre) und diskriminanter Validität (geringere Korrelationen mit Variablen, die von der Logik her nicht mit effektiver Lehre korrelieren sollten) anhand eines eigenen Lehrevaluationsinventars überprüft: Bedeutsam und konsistent korrelierten studentische Beurteilungen (student ratings) der Effektivität der Lehre mit den Beurteilungen früherer Studierender, den studentischen Leistungen (student achievement) in *multisection validity studies*, den Selbstevaluationen der Fakultäten hinsichtlich der eigenen Lehre und den Ergebnissen geschulter Beobachter bezüglich der Klarheit des Dozenten (teacher clarity). Gering korrelierten Forschungsleistung und Beurteilungen von Kollegen auf Basis von Visitationen des Unterrichts mit studentischen Lehrevaluationsergebnissen. Nach Überprüfung dieser Annahmen schlussfolgert Marsh, dass Konstruktvalidität gegeben sei.

Onwuegbuzie et al. (2007) fassen unter dem Begriff der *konstruktbezogenen Validität* verschiedene Aspekte nach Messick (1989b) zusammen - die laut Benson (1998, S. 11) in Teilen den dreien von Loevinger (1957) ähneln (siehe 6.2.1) - und zeigen auf, welche Befunde zu diesen existieren. Das gemessene Merkmal sei die *Wahrnehmung effektiver Eigenschaften von College-Dozenten* durch die Studierenden:

- *Substanzielle Validität (substantive*): Gibt an, ob die Art des studentischen Beurteilungsprozesses konsistent mit dem zu messenden Konstrukt ist. Evidenzen in dieser Hinsicht seien noch nicht erbracht.
- *Strukturelle Validität:* Beinhalte die Beurteilung, wie gut die Auswertungsstruktur (scoring structure) eines Instruments mit der Konstrukt-Domäne korrespondiert. Faktorenanalysen allein seien allerdings eine atheoretische Vorgehensweise. Evidenz solle primär in der Literatur anhand

von Vergleichen mit Items von Inventaren gesucht werden, bei denen relevante Eigenschaften gefunden wurden.

- *Ergebnis-Validität (outcome):* Die Bedeutung der Testwerte und die beabsichtigten und unbeabsichtigten Konsequenzen der Testnutzung. Hierbei würden Fragen gestellt werden, wie "Spiegelt der Inhalt des Lehrevaluationsinventars die Eigenschaften effektiver Unterrichtsmethoden wider, die von den Studierenden geschätzt werden?"
- *Generalisierbarkeit:* Das Ausmaß, in dem Bedeutung und Verwendung eines Instruments auf andere Populationen übertragbar ist. Bisherige Studien zeigten, dass sich Beurteilungen des Dozenten hinsichtlich der Fachrichtung und dem Niveau des Kurses (course level) unterscheiden. Es sei unklar, ob die Beziehung zwischen studentischen Urteilen und studentischer Leistung invariant ist und somit könne keine Aussage über eine Generalisierung getroffen werden.
- *Vergleichende Validität (comparative):* Hierzu zählen konvergente und diskriminante Validität, und die Autoren fassen hierfür verschiedene Studienergebnisse zusammen. Hinsichtlich konvergenter Validität zeigte sich Evidenz hinsichtlich Selbst-, Beobachter-, Kollegen- und Alumni-Beurteilungen und bezüglich diskriminanter Validität Selbstdarstellung (showmanship), Körpersprache, Milde bei Notenvergabe, Stimmlage und Gestik.

Die Autoren des Artikels schlussfolgern aus ihrer Sammlung von Ergebnissen, dass mehr Evidenz gebraucht werde, um einen Nachweis von Konstruktvalidität erbringen zu können.

Die hier dargelegte Studienlage spiegelt den Entwicklungsverlauf der Konstruktvalidität wider: Während die erste empirische Studie im Sinne des schwachen Programms anhand vielfältiger Korrelationen im Sinne konvergenter und diskriminanter Validität untersucht, versucht die zweite Arbeit verschiedene Quellen von Evidenz zusammenzuführen. Eine Reflektion der Konstruktvalidität im Kontext studentischer Lehrevaluation wird in Abschnitt 6.2.8 geführt. Zunächst werden die Ansätze dargestellt, die zeitlich nach der Konstruktvalidität aufgestellt wurden und im engeren Sinne als *argumentationsbasiert* anzusehen sind.

6.2.2 Interpretive Argument

Interpretation von Testwerten

Die *Interpretative Argumentation* (*Interpretive Argument*) nach Kane (1992) basiert auf der bereits erläuterten Annahme, dass sich Validität auf die Interpretationen der Testwerte bezieht und nicht auf die Werte oder den Test an sich. Dementsprechend seien diese Interpretationen zu evaluieren (Kane bezieht sich dabei auf die Standards von 1985: AERA, APA & NCME, 1985).

Mit dem Begriff der *Interpretation* seien „Bedeutung“ und „Erklärung“ assoziiert. Die Interpretation eines Testwertes impliziere somit, seine Bedeutung zu erklären und zumindest einige seiner Implikationen klar darzustellen. Die Validität einer Interpretation sei gegeben, wenn die Argumentation für ihren Nachweis plausibel ist.

Validierung der Interpretation

Die Testwerte sind der Ausgangspunkt dieser Argumentation, und die in der jeweiligen Testwert-Interpretation beinhalteten Aussagen und Entscheidungen repräsentieren die entsprechenden Schlussfolgerungen. Die *Schlüsse (inferences)* solch einer interpretativen Argumentation hingen von vielfältigen *Vorannahmen (assumptions)* ab, die mehr oder weniger glaubwürdig seien. Zum Beispiel hingen Schlüsse von einem Testergebnis auf ein Verhalten außerhalb eines Tests von Annahmen über deren Beziehung ab. Oder Schlüsse von Testwerten auf theoretische Konstrukte hängen von Annahmen in der Theorie ab, die das Konstrukt beinhalte.

Ein argumentationsbasierter Ansatz nutze dementsprechend die interpretative Argumentation als Rahmen für die Sammlung und Darstellung empirischer Evidenzen für diese Annahmen und Schlüsse. Die Validierung einer Testwert-Interpretation bedeute, die Plausibilität einer interpretativen Argumentation anhand angemessener Evidenzen zu unterstützen. Nicht alle Annahmen in der interpretativen Argumentation könnten bestätigt werden, aber es sollte anhand von Evidenzen gezeigt werden, dass sie hochgradig plausibel sind.

Eine interpretative Argumentation könnte zusammengefasst folgenden Verlauf annehmen:

1. Festlegung, welche Aussagen und Entscheidungen auf den Testwerten beruhen.

2. Spezifikation der Annahmen und Schlussfolgerungen, die von den Testwerten zu den Aussagen und Entscheidungen führen
3. Identifikation potentieller konkurrierender Interpretationen.
4. Suche nach Evidenz, die die Annahmen und Schlussfolgerungen der vorgeschlagenen interpretativen Argumentation unterstützen und potenzielle Gegenargumente zurückweist.

Praktische Argumentationen wie die der interpretativen Argumentation beinhalteten im Kontrast zu formalen Argumentationen Logik und Mathematik. Allerdings können nicht alle Annahmen und Schlussfolgerungen auf diese Weise evaluiert werden, sondern seien spezifisch für das jeweilige Themenfeld. Dieser Prozess müsse formalen Regeln folgen. Für die Evaluation praktischer Argumente sollten drei allgemeine Regeln beachtet werden:

1. Die Klarheit der Argumentation: Detaillierte Spezifikation der Schlüsse, Schlussfolgerungen und Annahmen, so dass klar ist, was die Argumentation beinhaltet.
2. Die Kohärenz der Argumentation: Die auf den Annahmen basierenden Schlussfolgerungen sind sinnvoll.
3. Die Plausibilität der Annahmen: Sind die Annahmen an sich plausibel oder durch Evidenzen gestützt, und inwieweit wirken sich schwache Annahmen auf die allgemeine Plausibilität aus?

In einem später erschienenen Artikel unterteilt Kane (2004) seinen argumentationsbasierten Ansatz in zwei Argumentationen: Einer interpretativen, in der die angenommenen Interpretationen und Verwendungen detailliert deutlich gemacht werden und einer Validitäts-Argumentation, in der die Kohärenz der interpretativen Argumentation und die Plausibilität ihrer Annahmen und Schlüsse evaluiert werden.

Hinsichtlich der Evaluation der interpretativen Argumentation würden laut Kane häufig folgende sechs Schlüsse zu berücksichtigen sein: Beobachtung, Generalisierung, Extrapolation, theoriebasierte Schlüsse, Entscheidungen und technische Schlüsse. In aktualisierter Form werden diese in Abschnitt 6.2.5 erläutert.

Beispiel

Porter (2011) bezieht sich bei einem Validitätsnachweis im Kontext des in den USA weit verbreiteten Fragebogens *National Survey of Student Engagement (NSSE)* auf den argumentationsbasierten Ansatz von Kane. Der Fragenbogen soll das Verhalten und die Einstellungen von Studierenden messen, zum Beispiel hinsichtlich ihrer Teilhabe an Universtäten und Colleges in Nordamerika.

Anhand von fünf Argumentationssträngen überprüfte er Annahmen der bisherigen NSSE-Forschung. Für die Überprüfung führte er keine eigene empirische Studie durch, sondern griff auf die bisherige Literatur zurück:

1. *Hintergrund:* Der NSSE sei spezifisch dafür entwickelt worden, das Ausmaß zu erfassen, in dem Studierende in empirisch abgeleiteten guten Bildungs-Maßnahmen eingebunden sind, und inwiefern sie von ihren College-Erfahrungen profitieren.
2. *Inhalt:* Der Fragebogen besteht aus Items, die sich direkt auf die institutionellen Beiträge hinsichtlich des Engagements der Studierenden, der College-Outcomes und der institutionellen Qualität beziehen.
3. *Antwort-Prozesse:* Die Items werden von Studierenden verstanden und korrekt beantwortet.
4. *Interne Struktur:* Die Items korrelieren derart miteinander, so dass sie in fünf Konstrukte gruppiert werden können: *Niveau der akademischen Herausforderung (Level of academic challenge), aktives und kollaborierendes Lernen, die Interaktion zwischen Studierenden und der Fakultät, bereichernde Bildungserfahrungen* und *eine unterstützende Campus-Umwelt*.
5. *Beziehung zu anderen Variablen:* Items und Skalen korrelieren mit anderen Daten (insbesondere mit Leistungs-Tests).

Porter kommt zu dem Schluss, dass die Validitäts-Argumentation bislang kein erwünschtes Ergebnis erbracht habe:

- Zu den Punkten 1 und 2 urteilte der Autor, dass die Spezifikation der Domäne des NSSE zu breit sei und statt durch theoretische Überlegungen anhand empirischer Ergebnisse zustande kam.
- Zu Punkt 3: College-Studierende hätten Probleme, Verhalten und Ereignisse zu berichten, insbesondere, wenn sie alltäglich seien. Daher

beruhten die Ergebnisse auf einer Reihe von Schätz-Strategien, die in Verzerrungen resultieren könnten.

- Zu Punkt 4: Die dimensionale Struktur wurde bislang in keiner Studie repliziert.
- Zu Punkt 5: Bisherige Forschung zeige, dass die Skalen mit keinerlei objektiven Maßen korrelierten.

6.2.3 Evidentiary Argument (Evidence Centered Design)

Ein weiterer argumentationsbasierter Ansatz ist das *Evidence Centered Design (ECD)*. Das ECD setzt einen Fokus auf die Entwicklung einer breiten Spanne von Assessment-Typen (unter anderem gewöhnliche Tests und Portfolios) und bietet einen allgemeinen, konzeptionellen Rahmen für verschiedene Elemente eines kohärenten Assessments. Mislevy, Almond & Lukas (2003); Mislevy, Steinberg & Almond (2003)

Durch diesen Rahmen werde gesichert, dass die Art, nach der Evidenz im Sinne der diagnostischen Ergebnisse gesammelt und interpretiert werde, konsistent mit dem Hintergrundwissen und Zwecken des intendierten Assessments sei. Er solle die Verbindungen zwischen dem Zweck, der Fertigkeits-Konzeption in einer Domäne, dem Design der Assessment-Elemente sowie der operationalen Prozesse zu verstehen helfen.

In einem Assessment würden Daten generiert. Aber das Interesse beziehe sich nicht auf diese, sondern auf die Anhaltspunkte (clues), die diese beinhalten. Diese seien Behauptungen (claims), die man über Personen in einem Assessment auf Basis von Beobachtungen mache. Die Natur und die Feinkörnigkeit (grainsize) einer Assessment-Annahme werde vom Zweck eines Assessments bestimmt. Die Aufgabe des Bestimmens der Relevanz von Assessment-Daten und ihres Wertes als Evidenz hinge von der Begründungskette von der Evidenz zu den Annahmen ab.

Grundlage des ECD sei das *Evidentiary Argument*: Das Evidentiary Argument beinhalte einen logischen Begründungszusammenhang zwischen den Ansprüchen oder Interpretationen und der Evidenz, um die Ansprüche zu unterstützen. In diesem Evidentiary Argument seien Konstrukt-Definitionen, Assessment-Aufgaben-Charakteristika und psychometrische Modelle für die Datenanalyse integriert. Ein Evidentiary Argument sei erforderlich, da Daten immer komplexer würden. Hierbei

könne eine Begründungskette von der Beobachtung bis zu den Schlüssen geführt werden.

Da dieser komplexe argumentationsbasierte Validitätsansatz nicht für das Thema dieser Arbeit relevant ist, wird auf die umfangreiche Literatur verwiesen. Beispiele zur Anwendung werden an verschiedenen Tests in Mislevy, Almond und Lukas (2003) dargestellt.

6.2.4 Assessment Use Argument

Beschreibung

Die *Assessment-Verwendungs-Argumentation (Assessment Use Argument, AUA*) wurde maßgeblich im Kontext von Sprachtests und der an diese gestellten Anforderungen entwickelt. Dabei wird insbesondere die Verbindung der Validität mit den Konsequenzen einer Testnutzung behandelt und nicht nur quantitative Messungen, sondern auch verbale oder visuelle Beschreibungen berücksichtigt. (Bachman, 2015)

Hintergrund dieses neuen Ansatzes war die Ansicht von Lyle Bachman (2003), den argumentationsbasierten Ansatz zur Validierung von Testwert-Interpretationen um eine Argumentation zu der Testverwendung (Use) zu erweitern: Nach der interpretativen Argumentation von Kane (siehe Abschnitt 6.2.2) sei eine Argumentation anhand unterstützender Evidenz bezüglich der jeweiligen Testwertinterpretation notwendig. Allerdings reiche nach Bachman (2005) diese Argumentation nicht aus, um die auf der entsprechenden Interpretation basierenden Verwendungen zu rechtfertigen. Es gebe keine Garantie dafür, dass valide Testwert-Interpretationen relevant, nützlich oder ausreichend für beabsichtigte Verwendungen seien. Die Interpretationen könnten auch untergraben (subverted) und für ursprünglich nicht intendierte Entscheidungen genutzt werden. Zusätzlich spricht sich Bachman (2015) für eine Trennung in auf den Testwert-Interpretationen basierenden *Entscheidungen* und den daraus resultierenden *Konsequenzen* aus (siehe Abbildung 12).

Validierung

Der Validierungsprozess beinhaltet wie bei der interpretativen Argumentation zwei Schritte (Bachman, 2015):

1. Die *Assessment Verwendungs-Argumentation* stellt einen konzeptuellen Rahmen für eine Argumentation dar, in der die beabsichtigten Verwendungen eines Assessments gerechtfertigt werden.
2. Im Anschluss werden in einer *Validitäts-Argumentation* Evidenzen gesammelt, die die *Assessment Verwendungs-Argumentation* stützen.

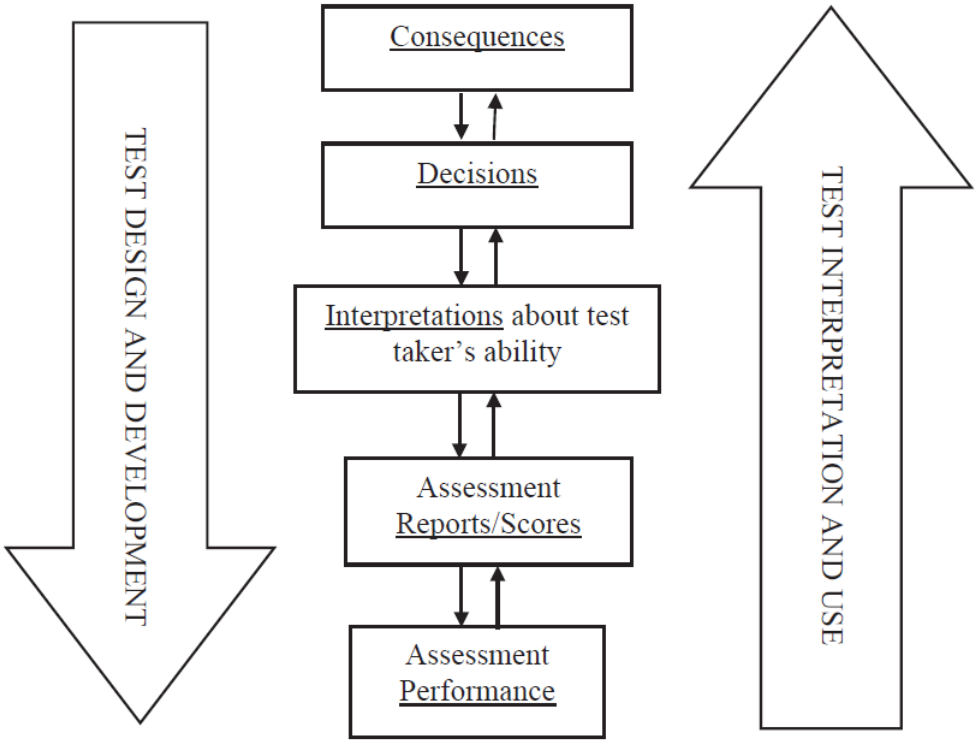

Abbildung 12: Die Rückschlüsse von den Konsequenzen eines Tests bis zum ursprünglichen Testverhalten nach Bachman & Palmer (2010, S. 91)

Die Verwendungs-Argumentation selbst beinhaltet eine *Assessment Validitäts-Argumentation* (*Assessment Validity Argument)* und eine *Assessment Nutzungs-Argumentation (Assessment Utilization Argument).* In ersterer wird das Testverhalten mit einer Interpretation verbunden und bei letzterer eine Interpretation mit einer Entscheidung. Somit sind in der AUA die Messung des Testverhaltens, die Interpretation der resultierenden Testwerte und auf der Interpretation basierende Entscheidungen in Bezug auf das gezeigte Verhalten berücksichtigt. Beide Argumentationen werden im Anschluss im Rahmen der Validitäts-Argumentation anhand von Evidenzen überprüft. (Bachman, 2005)

Beispiel

Die AUA wird maßgeblich bei Sprachtests verwandt. Dieser Ansatz wurde in einer Studie zu *British Sign Language (BSL)* anhand eines *Nonsense Sign Repetition Tests (NRST)* angewandt, um die Eignung des Tests zu überprüfen (Mann & Marshall, 2010): Das Verhaltens eines Probanden in diesem Test wird als valide Messung seines Phonologischen Arbeitsgedächtnisses interpretiert.

Auf der Ebene der *Assessment Validitäts-Argumentation* wird als Evidenz die große Menge bisheriger Studienergebnisse angeführt.

Bezüglich der Ebene der *Assessment Nutzungs-Argumentation* wurden vier Aussagen getroffen und überprüft:

1. *Das Assessment ist für die Messung der interessierenden Kompetenz relevant:* Als Evidenz wurde ebenfalls vorhandene Literatur angeführt. Allerdings könne eine konkurrierende Interpretation lauten, dass das visuell-räumliche Gedächtnis getestet werde, da es sich um eine Zeichensprache handele. Um dies zu überprüfen, wurden hörende Kinder, die keine Erfahrung in Zeichensprache hatten, mit dem *Nonsense Sign Repetition Test* getestet und diese schnitten bedeutsam schlechter als taube Kinder ab, woraus geschlossen wurde, dass diese die Aufgaben anders verarbeiteten.
2. *Interpretationen der NRST-Werte liefern Informationen, die für angemessene Entscheidungen genutzt werden können, da sie mit breiteren BSL-Fähigkeiten korrelieren:* Diese Aussage wurde bestätigt, da eine Korrelation des Tests mit dem *BSL Receptive Skills*-Test (auch bei Kontrolle des Alters) und auch eine Korrelation mit allgemeinen sprachlichen Fähigkeiten vorlag.
3. *Auf Basis des Assessments erhält die Testperson Unterstützung (zum Beispiel bei Mängeln in der Anwendung der Zeichensprache):* Die Literatur zeige, dass Anwender von Tests zu Sprachentwicklungsstörungen Ergebnisse nutzen und je nach Ergebnis spezifische Unterstützung bieten und anfordern.
4. *Ergebnisse des NSRT bieten ausreichende Informationen, um eine Entscheidung zu treffen:* Das gemessene Konstrukt ist bei der phonologischen Produktivität verortet. Entsprechend werden perzeptuelle und produktive Fähigkeiten gemessen, sowie die Fähigkeit erfasst, eine phonologische Repräsentation zur Speicherung im phonologischen Arbeitsgedächtnis zu enkodieren und von dort abzurufen. Weiterhin sind die Zeichen des Tests

keine realen lexikalischen Items und es werden auch keine morphologischen und syntaktischen Fähigkeiten erfasst.

Diese vier Aussagen unterstützen nach Aussage der Autoren die Entscheidungen von Anwendern, tauben Schülern eine bestimmte Art von Intervention zu bieten, wenn sie einen niedrigen Testwert erreichten. Die Autoren diskutieren auch Gegenargumente für diese Verwendung, auf die aus Platzgründen hier nicht eingegangen werden soll.

6.2.5 Interpretation/Use Argument

Beschreibung

Kane veröffentlichte 2013 (Kane, 2013) eine Weiterentwicklung der *interpretativen Argumentation* (Abschnitt 6.2.2) in Form der *Interpretations-/Verwendungs-Argumentation (Interpretation/Use-Argument, IUA)*.

In dieser werden die Validierungen von Interpretationen und von Verwendungen ebenbürtig berücksichtigt. Einem alleinigen Fokus auf die Verwendung in der *Assessment Verwendungs-Argumentation* (Abschnitt 6.2.4) begegnet Kane derart, dass Testwert-Interpretationen Annahmen (Claims) über Probanden beinhalteten und Testwert-Verwendungen Entscheidungen bezüglich der Probanden. Somit seien beide miteinander verschlungen und eine *Interpretations-/ Verwendungs-Argumentation* angemessen: Manche IUAs hätten einen Fokus auf eine bestimmte Verwendung, und andere beinhalteten eine Interpretation bezüglich einer Fähigkeit oder Verhaltensdisposition, die aber auch wiederum verschiedene Verwendungen erlaube.

Laut Kane sei die Validität einer angenommenen Interpretation oder der Verwendung eines Testwertes zu einem bestimmten Zeitpunkt durch die Plausibilität und Angemessenheit der angenommenen Interpretation oder Verwendung zu diesem Zeitpunkt definiert. Solch eine angenommene Interpretation oder Verwendung von Testwerten könne in dem Ausmaß als valide angesehen werden, in dem sie kohärent, vollständig und ihre Annahmen beziehungsweise Vorausetzungen entweder a priori plausibel sei oder durch Evidenzen gestützt werde (2013, S. 2-3).

Der Validierungsprozess

Der Ablauf eines entsprechenden Validitätsprozesses enthält folgende Schritte, die sich begrifflich in zwei Teile gliedern lassen, aber praktisch miteinander verwoben sind:

1. *Interpretations- /Verwendungs-Argumentation:* Die Argumentation wird spezifiziert und daraus resultierende Annahmen aufgestellt.
2. *Validitäts-Argumentation*: Diese beinhaltet die Evaluation aller Annahmen.

Anhand der *Interpretations- /Verwendungs-Argumentation* soll eine Interpretation oder Verwendung verständlich dargestellt und begründet werden. Die Argumentation sollte die angenommene Interpretation oder Verwendung reflektieren und nicht an eine vorgefertigte Struktur, sondern dem entsprechenden Kontext und der Population angepasst sein. Sie ist ein Netzwerk von Schlüssen (Inferences) und Vorannahmen (Assumptions), das vom Testverhalten zu Schlussfolgerungen und jeglichen auf diesen basierenden Entscheidungen führt. Dadurch wird explizit festgehalten, was angenommen wird und ein Rahmen für eine Validierung geboten.

Unter *Vorannahmen* wird beispielsweise verstanden, dass ein zu messendes Merkmal über die Zeit stabil ist. *Schlüsse* werden in einer „wenn-dann"-Form aufgestellt. Laut Kane kommen die folgenden Schlüsse in den meisten Argumentationen vor:

1. *Auswertung (Scoring):* Vom beobachteten Testverhalten zum beobachteten Testwert. Beispiel: Wenn ein beobachtetes Verhalten bestimmte Charakteristika enthält, dann erhält es einen bestimmten Testwert.
2. *Generalisierung:* Annahmen, wie sich das beobachtete Verhalten der Testperson über eine größere Anzahl von Aufgaben, verschiedenen Bedingungen und Gegebenheiten verändert, oder ob ein geschätzter Trait-Wert Schlussfolgerungen über zukünftiges Verhalten in einem Test zulässt (beispielsweise auf IRT-Modellen oder der Generalisierbarkeitstheorie basierend). Beispiel: Wenn ein hoher Wert bei einem Mathematiktest vorliegt, dann kann davon ausgegangen werden, dass die Person grundsätzlich mathematische Aufgaben lösen kann und auch im weiteren Verlauf in fortgeschrittenen Tests gut abschneidet.

3. *Extrapolation:* Wie der Testanwender sich in anderen Domänen, beziehungsweise wie er sich in der Zukunft in verschiedenen Arten von Aufgaben und Kontexten verhält (beispielsweise auf Erfahrungen oder Regressionsanalysen basierend). Beispiel: Wenn eine Person einen hohen Wert in einem Intelligenztest aufweist, dann lernt sie bestimmte technische Fertigkeiten schneller als andere.
4. *Kausal:* Erklärung des Testverhaltens durch Konstrukte (das Konstrukt entspreche der Interpretation). Beispiel: Wenn Personen bei verschiedenen Aufgaben zum Problemlösen jeweils ein ähnliches Muster aufweisen, dann kann das jeweilige Intelligenzniveau als Erklärung herangezogen werden.
5. *Entscheidung:* Die Annahmen über Konsequenzen verschiedener Entscheidungen für Personen mit verschiedenen Testwerten. Beispiel: Wenn eine Person bei einer Personalauswahl einen bestimmten Testwert erhält und auf eine bestimmte berufliche Position berufen wird, kann angenommen werden, dass sie die dortigen Aufgaben besser erfüllt als Personen mit einem niedrigeren Testwert.

Die *Validitäts-Argumentation* beinhaltet die Gesamt-Evaluation aller Ansprüche. Die Interpretationen/Verwendungen gelten in dem Ausmaß als valide, in dem

- eine IUA klar, kohärent und vollständig ist sowie ihre Schlussfolgerungen begründet sind. Es ist eine plausible Begründung für die angenommenen Interpretationen und Verwendungen gegeben, und keine essenziellen Schlüsse und Vorannahmen wurden ausgelassen.
- die Annahmen und Schlüsse an sich plausibel oder anhand angemessener Evidenzen gestützt sind.

Letztendlich sei solch ein Validierungsprozess nie als abgeschlossen anzusehen, da stets weitere Interpretationen und Argumente aufkommen können.

6.2.6 Standards for Educational and Psychological Testing

Validitätsbegriff

Auch in den *Standards for Educational and Psychological* Testing (AERA et al., 2014) wird ein Ansatz verfolgt, nach dem Validität ein einheitliches Konzept sei, es keine getrennten Validitätsarten gebe, und nicht der Test an sich valide sei. Wie schon unter 5.1.1 aufgeführt, wird Validität in den Standards derart definiert, so dass sie das

Ausmaß repräsentiere, in dem Theorie und Evidenz diejenigen Interpretationen von Testwerten unterstützen, die für die Verwendung des Tests vorgesehen seien. Sie repräsentiere den Grad aller akkumulierter Evidenz, die die intendierten Testwertinterpretationen für den angenommenen Nutzen unterstützt.

Validierungsprozess

Eine Validierung könne als ein Prozess angesehen werden, in dem Argumentationen konstruiert und evaluiert werden, die für oder gegen die intendierte Testwert-Interpretation und ihrer Relevanz für die angenommene Verwendung sprächen. Eine gut fundierte Validitäts-Argumentation integriere verschiedene Stränge von Evidenz in einer kohärenten Darstellung des Ausmaßes, in dem Evidenz und Theorie die beabsichtigte Interpretation von Testwerten für eine bestimmte Verwendung unterstützen. Hierbei könne sich herausstellen, dass die Konstruktdefinition überarbeitet, der Test verändert oder weitere Aspekte noch untersucht werden sollten.

Ein Validierungsprozess soll mit der Nennung der angenommenen Interpretation der Testwerte beginnen und dabei die Relevanz der Interpretation für die angenommene Verwendung begründen. Dabei wird das Konstrukt spezifiziert, das der Test messen soll, sowie die Konstrukt-Interpretation, die auf dem Antwortmuster basiert. Die beabsichtigte Konstrukt-Interpretation soll einmal anhand der Beschreibung ihres Geltungsbereichs (scope) und ihres Umfangs (extent) ausgearbeitet werden, sowie anhand der Darstellung der Aspekte des Konstrukts, die wiedergegeben werden sollen. Dies forme einen konzeptuellen Rahmen, mit dem das zu messende Konstrukt von anderen abgegrenzt werde und spezifiziere, wie es sich zu anderen Variablen verhalten solle.

Dieser konzeptuelle Rahmen zeige die zu sammelnde Evidenz auf, anhand derer die angenommene Interpretation evaluiert werden solle. Hierbei können Grundannahmen (Propositions) und Ansprüche (Claims) entwickelt werden, die die angenommene Interpretation unterstützten. Auch rivalisierende Hypothesen können aufgestellt werden, die die angenommene Interpretation in Frage stellen. Auch könne ein Test weniger als das angenommene Konstrukt messen (construct deficiency) oder durch andere Antwortprozesse beeinflusst werden (construct contamination).

Die Grundannahmen der jeweiligen Interpretation werden anhand empirischer Evidenz, Literatursuche und logischer Analyse evaluiert.

Wenn ein Testnutzer den Test in einem anderen Gebiet als das ursprünglich intendierte anwendet, obliegt es ihm, Evidenzen für den spezifischen Kontext zu identifizieren.

In der Regel sollten verschiedene Quellen zur Unterstützung der angenommenen Interpretation für eine bestimmte Verwendung herangezogen werden. Aber nicht jede Art von Evidenz sei für jeden Fall nötig. Evidenz sei für jede Grundannahme einer Testwertinterpretation in Bezug auf eine spezifische Verwendung heranzuziehen. Die Standards nennen fünf Quellen möglicher Evidenz:

1. *Evidenz auf Basis des Testinhalts (Evidence based on Test Content):* Analyse der Beziehung zwischen dem Inhalt und dem Konstrukt, das ein Test messen soll. Zu dem Inhalt zählen Themen, Formulierung und Itemformat sowie Vorgabe (administration) und Auswertung.
2. *Evidenz hinsichtlich von Antwortprozessen (Evidence regarding response processes):* Annahmen über kognitive Prozesse bei Testpersonen werden überprüft. Die Passung zwischen Konstrukt und dem Verhalten oder Antworten werden anhand theoretischer und praktischer Analysen überprüft.
3. *Evidenz basierend auf der internen Struktur (Evidence based on internal structure)*: Angabe des Ausmaßes, inwiefern die Beziehungen zwischen Testitems und Testkomponenten zu dem Konstrukt passen, auf dem die angenommenen Testwertinterpretationen basieren (zum Beispiel Eindimensionalität).
4. *Evidenz hinsichtlich der Beziehung zu anderen Variablen (Evidence regarding relations to other variables):* Die beabsichtigte Interpretation für eine bestimmte Verwendung impliziert, dass das Konstrukt zu anderen Variablen in Beziehung steht. Zum Beispiel zu bestimmten Kriterien, die durch das Testergebnis vorhergesagt werden, Beziehungen zu anderen Tests, die dasselbe oder ein anderes Merkmal messen oder eine Generalisierung auf ein Verhalten außerhalb der Testsituation.
5. *Evidenz für Validität und Konsequenzen des Testens (Evidence for validity and consequences of testing):* Evidenz für die Evaluation der Schlüssigkeit der

angenommenen Interpretationen für die beabsichtigte Verwendung. Einige Konsequenzen ergeben sich direkt aus der Interpretation von Testwerten und deren beabsichtigte Verwendungen. Andere Konsequenzen können jenseits der beabsichtigten Interpretation oder Verwendungen liegen.

Ein Validierungsprozess ende nie, aber letztendlich könne ein Urteil auf Basis der gesammelten Evidenzen gefällt werden.

Beispiel

Als ein Beispiel für die Kombination der *Interpretations- und Verwendungs-Argumentation* nach Kane und den Evidenzquellen nach den Standards kann der Fragebogen *Students' Mental Load and Mental Effort in Biology Education* angeführt werden: Er wird im Schulunterricht eingesetzt und die Ergebnisse, die Schüler hierbei erzielen, werden als Ausmaß von *psychischer Auslastung (Mental Load)* und *psychischer Anstrengung (Mental Effort)* bei der Bearbeitung von Aufgaben im Biologie-Unterricht interpretiert. Die Ergebnisse sollen als Kontrollvariablen in der Biologie-Bildungsforschung verwendet werden. Diese Interpretation und Verwendung wurde anhand von Evidenzquellen bezüglich des Inhalts (durch Expertenurteile), der internen Struktur (anhand von IRT-Modellen) und der Beziehung zu anderen Variablen (wie mit der Prüfungsleistung) überprüft. Der Autor kam zu dem Schluss, dass der Fragebogen in der Bildungsforschung zu Biologie genutzt werden könne, um die beiden Kontrollvariablen zu erfassen. (Krell, 2017)

6.2.7 Die Validitäts-Argumentation

Die Mehrheit der erwähnten Ansätze beziehen sich auf eine *Validitäts-Argumentation (Validity Argument)*, anhand derer eine Evaluation der *interpretativen, Assessment Verwendungs-* oder *Interpretations- /Verwendungs-*Argumentation durchgeführt wird.

Die Struktur einer Validitäts-Argumentation im Sinne von Toulmin (1958, zitiert nach Kane, 2013, S. 11-12) wurde in mehreren der Ansätze vorgeschlagen: Sie bietet ein allgemeines Rahmenmodell und Begrifflichkeiten zur Analyse vorläufiger Schlüsse (siehe Abbildung 13):

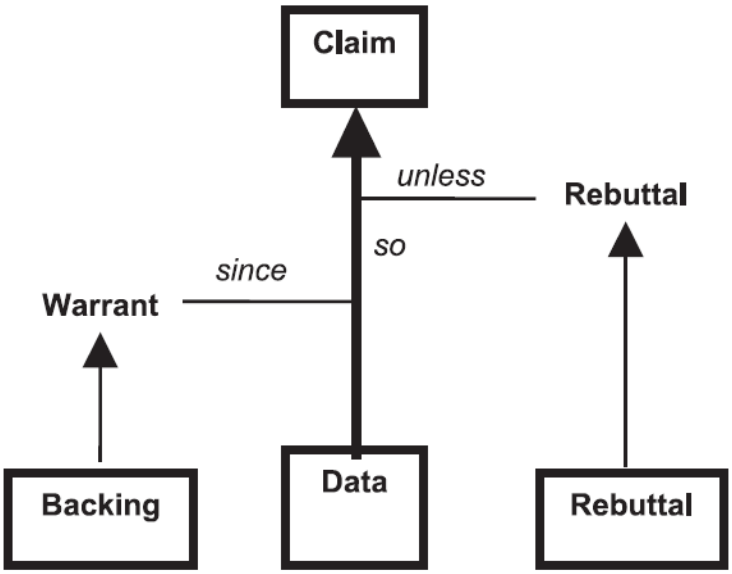

Abbildung 13: Diagramm einer Argumentationsstruktur nach Toulmin. Angepasste Version von Bachman (2015, S. 9). Ursprüngliche Version aus Mislevy, Steinberg et al. (2003, S.11).

Die einzelnen Aspekte dieser Argumentationsstruktur haben eine in Tabelle 3 beschriebene Bedeutung (Bachman, 2005; Kane, 2013). Hierbei wird eine Veranschaulichung von Bachman widergegeben, in dem ein aufgestellter Anspruch widerlegt wird.

Tabelle 3: Struktur einer Argumentation nach Toulmin mit Beispiel

Argumentationsaspekte	*Bedeutung*	*Beispiel*
Anspruch (Claim)	Der Anspruch entspricht bei argumentationsbasierten Ansätzen der Interpretation.	Marc ist ein US-Bürger.
Daten (Data)	Der Anspruch basiert auf Daten, die die entsprechenden Informationen beinhalten.	Marc wurde in den USA geboren.
Rechtfertigung (Warrant)	Schlussfolgerung von einem Anspruch oder Interpretation zu den Daten; beruht auf einer Rechtfertigung in Form einer allgemeinen Regel (bspw. eine Regressionsgleichung).	Alle Individuen, die in den USA geboren wurden, sind US-Bürger

Fortsetzung nächste Seite.

Fortsetzung Tabelle 3:

Unterstützung (Backing)	Die Rechtfertigung muss durch empirische Daten gestützt werden oder a-priori plausibel sein.	Nach der Verfassung ist jeder, der in den USA geboren wurde, ein US-Bürger
Widerlegung (rebuttal)	Da Ansprüche nicht immer gegeben sein können, können Ausnahmen angegeben werden (zum Beispiel für manche Populationen gilt aufgrund einer Behinderungen ein anderer Bezugsrahmen).	Marc hat auf seine US-Staatsbürgerschaft verzichtet
Daten im Sinne der Widerlegung (Rebuttal data)	Daten, die die alternative Erklärung unterstützen, schwächen oder ablehnen können	Marcs schriftliche Erklärung auf den Verzicht der US-Staatsbürgerschaft.
Qualifizierer (Qualifier)	Da die Rechtfertigung nicht allgemeingültig ist, sollte laut Kane (Kane, 2013) ein Qualifier eingesetzt werden, der die Stärke angibt (manchmal, gewöhnlich, fast immer). Statistisch spiegelt sich das in der Höhe von Standardfehlern oder Standardmessfehlern wider.	-

6.2.8 Argumentationsbasierte Ansätze und Konstruktvalidität

In dieser Arbeit wird Konstruktvalidität als Vorreiter der darauffolgenden argumentationsbasierten Ansätze beschrieben. Kane (2013, unter anderem auf S. 8) hat die Gemeinsamkeiten und Unterschiede zwischen beiden dargestellt. Kane bezog sich dabei auf die IUA, aber die genannten Aspekte können für alle weiteren argumentationsbasierten Ansätze ebenfalls als geltend angenommen werden. Die aus der folgenden Darstellung resultierende Schlussfolgerung wird als Ausgangspunkt für die Wahl des Validitätsansatzes für den empirischen Teil dieser Arbeit genutzt.

1. *Validierungsprozess:* Das einheitliche Modell der Konstruktvalidität wie das von Messick (1989b) beinhalte laut Kane keine Strategie, Validierungen durchzuführen. Im argumentationsbasierten Ansatz werde die Allgemeingültigkeit des einheitlichen Modells beibehalten und gleichzeitig ein direkter Ansatz für einen Validierungsprozess angeboten.
2. *Prinzipien der Konstruktvalidität werden beibehalten:* Die IUA übernehme die Rolle der formalen Theorie des starken Programms. Wenn die gemessene Eigenschaft ein theoretisches Konstrukt ist, soll die Theorie, die das Konstrukt definiert, den Kern der IUA ausmachen und den Ansatz von Cronbach und Meehl, (1955) imitieren.
3. *Schwaches Programm:* Die Endlosigkeit des schwachen Programms (in der jegliche Beziehung einer Eigenschaft berücksichtigt wird) und die Unklarheit (über die notwendige Menge von Evidenzen) als auch seine Mehrdeutigkeit (aufgrund nur vager Konstrukt-Interpretationen) können im argumentationsbasierten Ansatz vermieden werden. Im argumentationsbasierten Ansatz würden klare und vollständige Ansprüche aufgestellt und nur diese seien zu evaluieren.
4. *Starkes Programm:* Eine vollständig ausgearbeitete, formale Theorie sei nicht notwendig. Im argumentationsbasierten Ansatz brauchen nur Evidenzen für die Annahmen und die entsprechenden Voraussetzungen und Schlüsse erbracht werden.

6.3 Schlussfolgerungen

In dieser Arbeit wird der aktuellen Perspektive gefolgt, was unter Validität verstanden wird: Validität wird als einheitliches Konzept verstanden, welches sich graduell auf die Plausibilität der Interpretation und der auf ihr basierenden Nutzung von Testwerten bezieht.

Hinsichtlich von Lehrevaluationsergebnissen sind verschiedene Interpretationen und Verwendungen möglich, beziehungsweise als Ziel einer Datenerhebung angegeben (als Maß *effektiver Lehre*, *effektiver Dozenten*, von *Lernerfolg* oder *Zufriedenheit*), auf dem relevante Entscheidungen und Konsequenzen beruhen. Dementsprechend ist es sinnvoll, den Validierungsprozess im Sinne einer Argumentation zu verstehen, um für die angenommene Interpretation der Ergebnisse

eines bestimmten Lehrevaluationsinventars eine angemessene Grundlage vorzuweisen, und von anderen Interpretationen oder Verwendungszwecken plausibel abgrenzen zu können. Je nach zu evaluierendem Anspruch oder Schluss können verschiedene Analyseverfahren und auch herkömmliche Validitätsarten angemessen in eine Argumentation integriert werden.

Die beiden im Abschnitt 6.2.1 vorgestellten Artikel zum Thema Konstruktvalidität im Kontext studentischer Lehrevaluation zeigen auf, dass das schwache Programm auch hier in einer Ansammlung von Korrelationen mündet, und ein starkes Programm bislang nicht vollständig umgesetzt werden konnte. Nach Cronbach und Meehl (1955, S. 282) ist das starke Programm der Konstruktvalidität anzuwenden, wenn das zu messende Merkmal in einer Theorie eingebettet und nicht operational definiert ist. Eine operationale Definition beinhaltet, dass das zu messende Merkmal über Testinhalte definiert ist, und diese den interessierenden Bereich direkt repräsentieren (Hartig & Frey, 2007, S. 139-140). Allerdings sind die Grenzen zwischen einer rein operationalen Merkmalserfassung und dem Rückgriff auf eine Theorie fließend. Im Kontext studentischer Lehrevaluationen wird einerseits bei der Konstruktion von Inventaren eine inhaltlich theoretische Fundierung einbezogen, andererseits können dessen Ergebnisse nicht sinnvoll in einem nomologischen Netz überprüft werden, da eine starke Theorie weder aufgestellt wurde noch für die Validierung der Ergebnisinterpretation notwendig ist.

Aus den fünf anderen argumentationsbasierten Ansätzen wird für diese Arbeit der der *Standards of Educational and Psychological Testing* von 2014 ausgewählt (Abschnitt 6.2.6): Er repräsentiert den aktuellsten Ansatz und beinhaltet sowohl den Aspekt der Interpretation als auch der Verwendung. In der folgenden Argumentation wird dieser Ansatz durch die in den anderen Ansätzen vorgeschlagene Argumentationsstruktur nach Toulmin (Abschnitt 6.2.7) ergänzt, da diese eine formalisierte Darstellung und Evaluation der Grundannahmen gewährleistet.

7. Die Validität der Testwertinterpretationen studentischer Lehrevaluationsinventare und ihrer Verwendung

Im Folgenden soll der unter 6.2.6 skizzierte Validierungsprozess der *Standards for educational and psychological Testing* (AERA et al., 2014) angewandt werden:

1. Nennung der angenommenen Interpretation der Testwerte.
2. Darstellung, inwiefern sich diese von anderen Interpretationen abgrenzen.
3. Auf Basis dieses konzeptuellen Rahmens werden Grundannahmen für die Interpretation aufgestellt.
4. Überprüfung der Grundannahmen anhand (empirischer) Evidenzen.

7.1 Nennung der angenommenen Interpretation

Das Ziel des in dieser Arbeit berichteten Lehrevaluationsinventars ist die Erfassung von Lehrqualität im Sinne der *Zufriedenheit der Teilnehmer eines Unterstützungsprogramms für Promovierende mit der Durchführung der jeweiligen Lehrveranstaltung und der entsprechenden Vermittlung von Lehrinhalten*.

Im Sinne einer Interpretation würden höhere Ergebnisse für ein größeres Ausmaß der Lehrqualität im Sinne der genannten Zufriedenheit stehen.

Diese Interpretation gilt nach den *Standards* als valide, wenn theoretische Überlegungen und empirische Analysen diese Deutung unterstützen und somit praktische Konsequenzen aus den Ergebnissen gezogen werden können.

7.2 Abgrenzung zu anderen Interpretationen

In der in den vorangegangenen Abschnitten berichteten Literatur wurden verschiedene Konstrukte genannt, die anhand studentischer Evaluation gemessen wurden. Darunter fallen folgende Konstrukte:

1. *Studentische Lehrevaluationsergebnisse geben ein Ausmaß von Lernerfolg an:* Lernerfolg ist als maßgeblich zu messendes Konstrukt in dem hier verwandten Lehrevaluationsinventar nicht vorgesehen. Spezifisch auf das Promotionskolleg ausgerichtet, ist es schwierig, die Angaben von Promovierenden nach einer Lehrveranstaltung als Erfolg des Lernens ohne Leistungsüberprüfung messen zu wollen. Eine Promotion ist ein komplexer Prozess, zu dem die Veranstaltungen des Promotionskollegs durch die

Vermittlung grundlegender Kompetenzen wissenschaftlichen Arbeitens nur einen Teil beitragen.

2. *Studentische Lehrevaluationsergebnisse geben eine subjektive Einschätzung des Grads an Lehrkompetenz eines Dozenten durch die Studierenden wieder (effektiver Dozent):* Die Kompetenz des Dozenten ist ein elementarer Bestandteil hinsichtlich der Durchführung einer Lehrveranstaltung und der Vermittlung von Lehrinhalten. Allerdings besteht die Zufriedenheit mit solch einer Veranstaltung aus mehr als nur dem Dozenten: Unter anderem kommen noch die grundsätzliche Relevanz des Themas für die Teilnehmer, die Rahmenbedingungen und die Organisation unabhängig von der konkreten Lehrtätigkeit hinzu.

7.3 Grundannahmen im Sinne der Interpretation und entsprechende Evidenzen

Die in dieser Arbeit angenommene und beabsichtigte Interpretation studentischer Lehrevaluationsergebnisse wird mit mehreren entsprechenden Grundannahmen assoziiert. Diese werden im Folgenden dargestellt.

Für diese Darstellung wird das von den verschiedenen argumentationsbasierten Validitätsansätzen genannte Rahmenmodell für eine Argumentationsstruktur nach Toulmin verwandt (siehe Abschnitt 6.2.7), auch wenn es in den *Standards* nicht aufgeführt wird. Das Rahmenmodell wird für diese Arbeit als hilfreich angesehen, da hierdurch der Zusammenhang von der Interpretation über die Grundannahmen bis zu den entsprechenden Evidenzen strukturiert dargestellt werden kann.

Wie aus Tabelle 4 ersichtlich, werden in dieser Arbeit zwei Grundannahmen im Sinne der genannten Interpretation überprüft:

1. Es wird als grundlegend angenommen, dass alle qualitätsrelevanten Aspekte in einem Inventar aufgeführt sind, wenn die qualitätsbezogene Zufriedenheit der Studierenden mit der Durchführung der Lehrveranstaltungen eines Unterstützungsprogramms für Promovierende und der Vermittlung von Lehrinhalten erfasst werden soll. Dies entspricht der Inhaltsvalidität des klassischen Ansatzes.

2. Weiterhin wird als grundlegend angenommen, dass die den qualitätsrelevanten Inhalten entsprechenden Items die interessierenden Qualitätsunterschiede angemessen abbilden können.

Diese beiden Grundannahmen werden im weiteren Verlauf hinsichtlich der Interpretation der Evaluationsergebnisse überprüft. Aufgrund des Umfangs dieser Arbeit werden keine empirischen Überprüfungen weiterer Grundannahmen der Interpretation und für Verwendungen der Ergebnisse sowie auf den Verwendungen basierende Konsequenzen durchgeführt, aber in Abschnitt 9 skizziert.

Tabelle 4: Struktur einer Argumentation nach Toulmin auf studentische Lehrevaluation übertragen

Anspruch	Wenn Lehrevaluationsergebnisse für Lehrqualität im Sinne einer Zufriedenheit mit einer Lehrveranstaltung stehen, dann ….	
Daten	…, dann sind alle qualitätsrelevanten Inhalte vorhanden	…, dann differenzieren Items plausibel hinsichtlich ihres Iteminhalts
Rechtfertigung	Theoretische und empirische Fundierung	Varianzkomponentenschätzung anhand eines linearen Mischmodells
Unterstützung	Abgleich mit • Theorien • Studierendenbefragungen • Expertenbefragung • weiteren Studienergebnisse	Daten der studentischen Lehrevaluation im Promotionskolleg
Qualifizierer	Alle relevanten Inhalte	Plausible, angemessene Differenzierung
Widerlegung	-	-
Daten im Sinne der Wider-legung	-	-

7.4 Datengrundlage dieser Arbeit

Das *Frankfurter Promotionskolleg am Fachbereich Medizin* ist ein im Jahre 2011 etabliertes Unterstützungsprogramm für Promovierende (Sennekamp et al., 2016). Es hat zum Ziel, grundsätzliche Kompetenzen wissenschaftlichen Arbeitens zu vermitteln. Somit soll Promovierenden der Promotionsprozess erleichtert, und Betreuer sollen entlastet werden. Das Kolleg ist für Angehörige des Fachbereichs kostenlos und wird von Promovierenden auf freiwilliger Basis besucht. Es besteht aus acht sogenannten Grundkursen, die mehrmals im Jahr angeboten werden und jeweils zwei Stunden dauern: Die acht Kurse lauten „Gute wissenschaftliche Praxis“, „Literaturrecherche“, „Literaturverwaltung mit wahlweise Citavi oder Endnote“,

„Gliederung und Aufbau einer Dissertation“, „Textformatierung mit Word“, „Klinische Epidemiologie“ in zwei Kursen und „Datenmanagement mit Excel“. Teilnehmer können die Anzahl und den Zeitpunkt der Kurse selbst wählen. Jeder Kurs wird direkt im Anschluss von den Studierenden anhand eines einheitlichen Bogens evaluiert (siehe Anhang A). Die Teilnehmer werden gebeten, einen Code auf jedem Bogen zu hinterlassen, der sie als Person anonymisiert, aber die Bögen über verschiedene Kurse hinweg als zueinander gehörend identifizieren lässt. Über die Grundlagenkurse hinaus werden noch fakultative Kurse angeboten, die als spezifisches Angebot hinsichtlich bestimmter Promotionsthemen angesehen werden (statistische Verfahren, Fragebogenkonstruktion oder auch Präsentationstechniken). In dieser Dissertation werden ausschließlich die Grundlagenkurse berücksichtigt.

Die Lehrevaluationsergebnisse liegen in einer kreuz-klassifizierten Datenstrukturen vor: Die Teilnehmer können sich die unterschiedlichen Kurse nach ihrer eigenen Zeitplanung aussuchen und die Art und Anzahl der Kurse selbst bestimmen, die sie besuchen wollen. Somit gibt es Teilnehmer, die nur einen oder alle acht Kurse besuchten. Weiterhin gibt es Dozenten, die nur wenige Male Kurse unterrichteten und andere, die seit mehreren Jahren teilnehmen. Manche gaben nur zu einem oder zwei Themen Unterreicht, andere zu mehreren.

8. Die Überprüfung der Grundannahmen

Hinsichtlich der genannten Interpretation studentischer Lehrevaluationsergebnisse im Sinne einer qualitätsbezogenen Zufriedenheit mit einer Lehreveranstaltung werden im folgenden Verlauf zwei ihrer Grundannahmen überprüft.

8.1 Grundannahme 1: Alle qualitätsrelevanten Aspekte werden erfasst

Die erste Grundannahme beinhaltet die Anforderung, dass alle Aspekte in dem Inventar berücksichtigt werden, die für die Messung von Lehrqualität im genannten Sinne relevant sind.

Um dies zu gewährleisten, sollte ein Inventar nach wissenschaftlichen Kriterien gestaltet werden. Laut Rindermann ist für eine Konstruktion eines Inventars eine Mischung verschiedener Verfahren hilfreich (Rindermann, 2009, S. 59): Es können Ergebnisse der Lehrevaluationsforschung und der Instruktionspsychologie herangezogenen werden (siehe Kapitel 3), Items anderer Inventare übernommen und miteinander kombiniert, oder Studierende und Dozenten nach relevanten Kriterien befragt werden.

Das in dieser Arbeit verwendete Inventar wurde zum Beginn des Promotionskollegs im Jahre 2011 entwickelt. Zunächst wird daher die damalige Konstruktionsstrategie zusammengefasst, und im Anschluss die Überprüfung der Inhalte im Sinne der Grundannahme berichtet. Da es bei der Erfassung durch das Inventar um die Zufriedenheit der Teilnehmer geht, werden bei der Überprüfung zunächst die Ergebnisse von Studierendenbefragungen zu guter Lehre herangezogen und mit den Inhalten des Inventars verglichen. Diese Studierendenbefragungen beziehen sich allerdings auf allgemeine Lehrveranstaltungen. Da es sich beim Promotionskolleg aber um ein Unterstützungsprogramm für Promovierende handelt, wurden zusätzlich die Teilnehmer des Promotionskollegs befragt, welche bislang erhobenen Qualitätsaspekte sie für solch eine Veranstaltung als relevant einstufen, und welche sie darüber hinaus zusätzlich als relevant ansehen.

Auf Basis dieser beiden Evidenzquellen ist somit die Perspektive der Teilnehmer hinsichtlich ihrer eigenen Zufriedenheit abgedeckt. Allerdings ist es möglich, dass Teilnehmer von Lehrveranstaltungen nicht über einen umfassenden Überblick hinsichtlich der Vermittlung von Lehrinhalten verfügen. Aufgrund dessen werden die

Inhalte des Inventars durch zwei Expertinnen des Fachgebiets der Hochschuldidaktik auf ihre Relevanz hinsichtlich der Erfassung von Lehrqualität überprüft. Weiterhin wird das Inventar mit den Merkmalen bekannter Lehr-Lern-Theorien sowie mit den Inhalten eines allgemeinen Inventars und einem für ein weiteres Lehrangebot für Promovierende verglichen. Die daraus resultierenden Abweichungen werden dann hinsichtlich ihrer Relevanz für das hier untersuchte Inventar diskutiert.

8.1.1 Die ursprüngliche Konstruktion des Inventars

Bei der Konstruktion des Inventars wurde sich an dem Leitfaden des Kapitels „Planung und Entwicklung von psychologischen Tests und Fragebogen“ aus dem Buch „Testtheorie und Fragebogenkonstruktion“ orientiert (Jonkisz & Moosbrugger, 2007):

1. Zu Beginn sollte festgelegt werden, welches Konstrukt an welcher Zielgruppe in welchem Geltungsbereich mit welcher Art von Test zu messen ist. In dieser Arbeit ist das Konstrukt die Lehrqualität im Sinne der *Zufriedenheit der Teilnehmer eines Unterstützungsprogramms für Promovierende mit der Durchführung der jeweiligen Lehrveranstaltung und der entsprechenden Vermittlung von Lehrinhalten*. Inwiefern diese Zufriedenheit erfüllt wird, wird anhand eines Lehrevaluationsinventars erfasst (siehe Anhang A).
2. Im nächsten Schritt erfolgte die Generierung der Testaufgaben, die hier nur kurz angerissen werden soll. Mehrere Strategien stehen hierzu zur Verfügung, die auch miteinander kombiniert werden können (intuitive, rationale, externale und internale Konstruktionsstrategie). Zunächst wurde die *rationale Konstruktionsstrategie* genutzt, bei der aus vorhandenen Theorien über das zu messende Konstrukt Items abgeleitet werden. Hierfür wurden die Inhalte verschiedener Lehrbücher und bereits vorhandener Inventare herangezogen (resultierte in Items wie „Die Veranstaltung war gut organisiert“, „Die verwendeten Unterrichtsmaterialeien waren angemessen“ und „Ich hatte ausreichend Vorkenntnisse, um der Veranstaltung zu folgen“). Ebenso wurden auch Items *intuitiv* auf den spezifischen Anwendungsbereich hin entwickelt (wie „Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können“).
3. Im Anschluss wurde der Aufgabenstamm (Item als Aussage, Frage oder Aufgabe) und das entsprechende Antwortformat (frei, gebunden oder atypisch)

ausgewählt: Elf Items des Inventars wurden als Aussage mit vierstufiger Ratingskala ausformuliert („stimme nicht zu“, „stimme eher nicht zu“, „stimme eher zu“ bis „stimme zu“). Diese Abstufungen wurden als ausreichend zur Erfassung der Beurteilung von Qualitätsunterschieden aus studentischer Sicht angesehen. Um den Teilnehmern eine zusammenfassende Beurteilung zu ermöglichen, wurde eine Aussage in Form einer Schulnote eingefügt. Neben diesen wurden auch drei weitere Items mit dreistufigem Antwortformat (beispielweise ob der Zeitrahmen der Veranstaltung „zu kurz“, „genau richtig“ oder „zu lang“ war) und fünf mit freiem Antwortformat erstellt (zum Beispiel „Das fand ich besonders gut“).

Weiterhin werden soziodemographische Daten sowie der Status des eigenen Promotionsprozesses abgefragt.

Das aus diesem Konstruktionsprozess resultierende Evaluationsinventar ist seit Januar 2011 im Einsatz und wird in dieser Arbeit anhand folgender Evidenzquellen hinsichtlich seines Inhalts überprüft.

8.1.2 Werden aus Sicht der Teilnehmer alle qualitätsrelevanten Aspekte erfasst?

Zwei Studien (siehe Abschnitt 3.3) untersuchten, welche Kriterien Studierende mit guter Lehre in Verbindung bringen. Der Vergleich dieser Kriterien mit den Inhalten des Inventars wird in Tabelle 5 dargestellt.

Tabelle 5: Vergleich der Items des Inventars (PK) mit Ergebnissen von Studierendenbefragungen

Feldmann (1976)	*PK*	*Onwuegbuzie (2007)*	*PK*
Anregung (stimulation)	Die Durchführung durch den Dozenten war motivierend	Eingehend (Dozent gibt Rückmeldung über Leistung)	*Nicht in Inventar*
Klarheit und Verständlichkeit	Die Lernziele der Veranstaltung waren klar erkennbar Die Erklärungen des Dozenten waren verständlich	Begeistert (Dozent hat beispielsweise Spaß an der Lehre)	*Nicht in Inventar*
Das Wissen des Dozenten über den Unterrichtsgegenstand	*Nicht in Inventar*	Auf die Studierenden zentriert (zum Beispiel auf Probleme von Studierenden eingehend und starke zwischenmenschliche Kompetenzen besitzend)	*Nicht in Inventar*
Die Vorbereitung des Dozenten und die Organisation des Kurses	Die Veranstaltung war gut organisiert *Vorbereitung des Dozenten fehlt*	Professionell (zum Beispiel wurden Ziele gesetzt, die zu erreichen sind)	Die Stoffmenge der Veranstaltung war genau richtig Der Zeitrahmen der Veranstaltung war genau richtig Das Unterrichtstempo des Dozenten war genau richtig
Enthusiasmus des Dozenten für das Thema und für die Lehre	*Nicht in Inventar*	Experte (der Dozent besitzt Wissen über den Kursinhalt und darüber hinaus)	*Nicht in Inventar*
Freundlichkeit des Dozenten, Sorge und Respekt für Studierende	*Nicht in Inventar*	Verbindend (Der Dozent ist für Studierende erreichbar und kann dadurch zusätzliche Hilfe anbieten)	*Nicht in Inventar*

Fortsetzung nächste Seite.

Fortsetzung von Tabelle 5:

Verfügbarkeit und Hilfsbereitschaft des Dozenten	*Nicht in Inventar*	Übermittler/Vermittler (kann das Interesse des Kurses aufrechterhalten, hat gute sprachliche Qualitäten)	Die Erklärungen des Dozenten waren verständlich Die Arbeitsatmosphäre war konstruktiv Die Durchführung durch den Dozenten war motivierend
Ermunterung zu Fragen und Diskussionen, Offenheit für die Meinung von anderen	Die Arbeitsatmosphäre war konstruktiv Ich hatte die Möglichkeit, mich aktiv an der Veranstaltung zu beteiligen	Moralisch (zum Beispiel, werden vom Dozenten alle Studierenden gleich behandelt)	*Nicht in Inventar*
-	-	Leiter (Bietet eine sichere und geordnete Lernumgebung durch effiziente Zeitstruktur und optimierte Ressourcen)	Die Veranstaltung war gut organisiert

Bei Feldmann gibt es bei fünf Aspekten bezüglich des Dozenten keine Entsprechung im Inventar: Sein Wissen in Bezug auf den Unterrichtsgegenstand, seine Vorbereitung auf den Kurs, sein Enthusiasmus, seine sozialen Kompetenzen in Form von Freundlichkeit und Respekt sowie seine Verfügbarkeit und Hilfsbereitschaft.

Bei der zweiten Studie zeigen sich sechs nicht in dem Inventar vorkommende Aspekte hinsichtlich des Dozenten: Dass er Rückmeldung über die Leistung gibt, vom Unterrichten begeistert ist, Aspekte der zwischenmenschlichen Beziehung, sein Fachwissen, seine Erreichbarkeit und Hilfsbereitschaft sowie moralische Eigenschaften.

8.1.3 Welche qualitätsrelevanten Aspekte sind in einem Programm für Promovierende relevant?

Um diese Studienlage zu ergänzen und auf ein Programm für Promovierende auszurichten, wurde für diese Arbeit unter den Teilnehmern des Promotionskollegs eine Umfrage durchgeführt. Dies wurde als sinnvoll erachtet, da es sich um ein spezifisches Programm handelt, und es unter Umständen weitere Qualitätskriterien in einer Evaluation erfordert. Hierfür wurden die Items des aktuellen Inventars in einer Online-Umfrage von den Promotionskollegsteilnehmern hinsichtlich ihrer Relevanz bewertet.

Hierbei wurden die in den Items erfassten qualitätsrelevanten Aspekte von den Teilnehmern anhand einer vierstufigen Ratingskala („relevant“, „eher relevant“, „eher nicht relevant“ und „nicht relevant“) bewertet. Ausgeschlossen wurden das Item „Ich würde die Veranstaltung weiterempfehlen“ und die Gesamtbeurteilung in Form einer Schulnote. Diese erfassen nicht konkret qualitätsrelevante Aspekte, sondern spiegeln eine allgemeine Zufriedenheit wider.

Weiterhin wurde anhand einer Ja/Nein-Antwort erfragt, ob durch die genannten Aspekte alle qualitätsrelevanten Bereiche abgedeckt werden. Falls mit „Nein“ genantwortet wurde, konnten die Teilnehmer der Umfrage die ihrer Ansicht nach fehlenden Aspekte nennen.

Es wurden 294 Teilnehmer angeschrieben, von denen 79 Personen im Zeitraum vom 14.06.2017 bis 10.7.2017 den Fragebogen auf *Limesurvey* (Schmitz, 2012) beantwortet haben.

Wie aus Abbildung 14 ersichtlich, wurden 11 der 14 abgefragten Inhalte jeweils von über 90% der Teilnehmer als „relevant“ oder „eher relevant“ eingestuft. Zustimmung von 100% zu den beiden positiven Antwortmöglichkeiten gab es bei dem Item zur „Verständlichkeit des Dozenten“ und dem spezifischen Item, die Inhalte in der eigenen Dissertation umsetzen zu können. Die höchsten Anteile der Zustimmung zu „eher nicht relevant“ und „nicht relevant“, zeigten sich bezüglich der eigenen ausreichenden Vorkenntnisse, um der Veranstaltung folgen zu können (32.9%) und der Möglichkeit, sich aktiv an der Veranstaltung zu beteiligen (21.5%). Fehlende Werte gab es bei keiner Antwort.

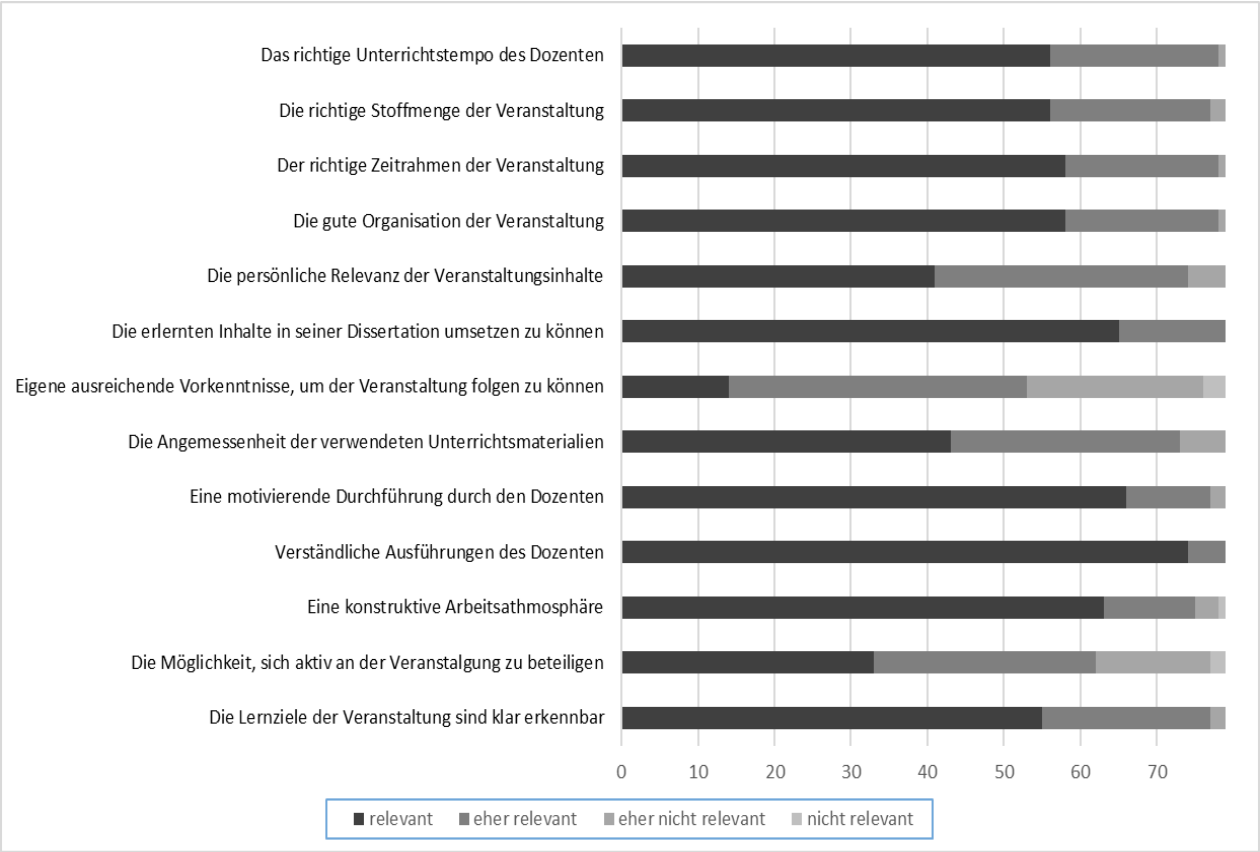

Abbildung 14: Ergebnisse der Studierendenbefragung hinsichtlich ihrer Einschätzung der Relevanz qualitätsrelevanter Aspekte in absoluten Zahlen (79 Teilnehmer)

68 der Teilnehmer (86.076%) hatten angegeben, dass durch die genannten Inhalte für sie alle qualitätsrelevanten Bereiche abgedeckt worden seien. Die weiteren elf hatten nicht in dem Inventar berücksichtigte Anmerkungen, die wie folgt zusammengefasst werden können:

- die Erfahrung des Dozenten
- die richtige Gruppengröße
- Durchführung und angemessener Zeitrahmen von Übungsaufgaben
- die Möglichkeit, auch nach Ende der Lehrveranstaltung Kontakt zum Dozenten aufnehmen zu können.

Von den Befragten wurden keine spezifischen Inhalte angegeben, die sich auf ein Lehrangebot zu Promotionen beziehen. Weitere zwölf der von den Teilnehmern genannte Aspekte sind schon in einer Form in dem Inventar vorhanden (zum Beispiel

„genügend Zeit“ als Aussage zum Zeitrahmen oder „Handouts“ als Aussage zu der Angemessenheit der Unterrichtsmaterialien).

8.1.4 Übereinstimmung des Inventarinhalts mit Theorien, Modellen und Studienergebnissen

In Abschnitt 3.2 wurden zwei etablierte Lehr-Lern-Theorien beziehungsweise Modelle erläutert. Diese können eine Grundlage für jede Lehrveranstaltung bieten (auch wenn das Angebots-Nutzungs-Modell von Helmke auf Schulen bezogen ist) und somit in einer Evaluation abzufragende Inhalte beschreiben. Dieser Abschnitt bezieht sich nur auf diese beiden Modelle. Die Ausarbeitungen zu guter Lehre wie der der Universität Mainz oder des Stanford Faculty Programms werden hier nicht berücksichtigt.

Tabelle 6: Vergleich der Items des Inventars (PK) mit denen als relevant identifizierten Aspekte des Angebots-Nutzungs-Modells und des Constructive Alignments

Helmke (2006)	*PK*	*Biggs (1996)*	*PK*
Effiziente Klassenführung und Zeitnutzung	Zeitrahmen Stoffmenge Unterrichtstempo	Klarheit der Lehrenden über Lernziele (Outcome)	Lernziele klar
Lernförderliches Unterrichtsklima	Die Arbeitsatmosphäre war konstruktiv	Auf die Lernziele ausgerichtetes Lehr-Lernverhalten	*Nicht in Inventar*
Vielfältige Motivierung	Die Durchführung durch den Dozenten war motivierend	Angemessene Übung	*Nicht in Inventar*
Strukturiertheit und Klarheit	Die Lernziele der Veranstaltung waren klar erkennbar Die Erklärungen des Dozenten waren verständlich	Evidenz für Übereinstimmung mit Zielen	*Nicht in Inventar*
Wirkungs- und Kompetenzorientierung	Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können Die Inhalte der heutigen Veranstaltung sind für mich persönlich relevant	-	-
Schülerorientierung	*Nicht in Inventar*	-	-

Fortsetzung folgende Seite.

Fortsetzung von Tabelle 6:

Förderung aktiven, selbstständigen Lernens	Ich hatte die Möglichkeit, mich aktiv an der Veranstaltung zu beteiligen	-	-
Angemessene Variation von Methoden und Sozialformen	Die verwendeten Unterrichtsmaterialen waren angemessen	-	-
Konsolidierung, Sicherung, intelligentes Üben	*Nicht in Inventar*	-	-
Passung	Ich hatte ausreichend Vorkenntnisse, um der Veranstaltung zu folgen Die Inhalte der heutigen Veranstaltung sind für mich persönlich relevant	-	-

Im Vergleich mit dem Angebots-Nutzungsmodell von Helmke zeigt sich eine großteilige Übereinstimmung zwischen den abgefragten Inhalten und dem Inventar (siehe Tabelle 6).

Was in beiden Modellen angegeben, aber nicht in dem Inventar konkret abgefragt wurde, ist der Aspekt der Konsolidierung und Übung des Erlernten.

Das Grundprinzip des Constructive Alignments beinhaltet die Bezugnahme der einzelnen Aspekte aufeinander. Dieses Grundprinzip könnte auch in ein Item überführt werden: „War ein roter Faden von den Lernzielen zu den gelehrten Inhalten und Übungen erkennbar?“ Da es keine Abschlussklausur oder ähnliches gibt, würde dieser Aspekt nicht Teil des roten Fadens werden.

8.1.5 Welche Kriterien werden in anderen Lehrevaluationsinventaren abgefragt?

Eine weitere Quelle liegt in schon bereits entwickelten Inventaren, die im praktischen Einsatz sind: Anhand dieser kann ebenfalls verglichen werden, ob relevante Aspekte noch zusätzlich für das eigene Inventar in Betracht kommen könnten. Für diese Arbeit wurden der allgemeine Teil des Lehrevaluationsinventares der Johann Wolfgang Goethe-Universität in Frankfurt am Main (Johann Wolfgang Goethe-Universität, 2011) herangezogen (siehe Anhang B), sowie eines weiteren Unterstützungsprogramms für Promovierende, dem der Goethe Graduate Academy (GRADE) (Goethe Graduate Academy, 2013); siehe Anhang C. Somit kann ein

Vergleich mit allgemeinen sowie mit spezifischen Qualitätskriterien aufgestellt werden (Tabelle 7).

Tabelle 7: Vergleich der Items des Inventars (PK) mit den Inventar-Inhalten

GRADE	*PK*	*Goethe-Universität*	*PK*
Der Workshop hat meine Erwartungen erfüllt.	Gesamtnote	Der Besuch der Veranstaltung führt zu einem spürbaren Wissenszuwachs	Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können
Der Workshop war klar strukturiert.	Die Lernziele der Veranstaltung waren klar erkennbar Die Veranstaltung war gut organisiert Der Zeitrahmen der Veranstaltung war genau richtig Die Stoffmenge der Veranstaltung war genau richtig	Der in der Veranstaltung vermittelte Stoff ist gut strukturiert	Die Lernziele der Veranstaltung waren klar erkennbar Der Zeitrahmen der Veranstaltung war genau richtig Die Stoffmenge der Veranstaltung war genau richtig
Der/die Trainer/in hat die Inhalte klar und anschaulich vermittelt.	Die Erklärungen des Dozenten waren verständlich Die verwendeten Unterrichtsmaterialien waren angemessen	In der Veranstaltung werden ausreichend Hilfsmittel zur Aneignung des Lernstoffs (Skripte, Lehrtexte, Literaturlisten etc.) angeboten	Die verwendeten Unterrichtsmaterialien waren angemessen
Ich würde den Workshop weiterempfehlen.	Ich würde diese Veranstaltung anderen Doktoranden weiterempfehlen	Das Tempo der Veranstaltung ist angemessen	Der Zeitrahmen der Veranstaltung war genau richtig
Der Inhalt des Workshops war hilfreich.	Die Inhalte der heutigen Veranstaltung sind für mich persönlich relevant Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können	Der Veranstalter / die Veranstalterin geht auf Fragen der Teilnehmer/-innen angemessen ein	*Nicht in Inventar*
Die Länge des Workshops war angemessen.	Der Zeitrahmen der Veranstaltung war genau richtig	In der Veranstaltung wird ein guter Überblick über das behandelte Stoffgebiet vermittelt	Stoffmenge Relevanz
Welche Workshopinhalte haben Sie besonders angesprochen?	Offenes Antwortformat: Das fand ich besonders gut	In der Veranstaltung sind inhaltliche Zusammenhänge („roter Faden“) deutlich erkennbar	*Nicht in Inventar*

Fortsetzung folgende Seite.

Fortsetzung von Tabelle 7:

Fehlten in diesem Workshop bestimmte Inhalte oder Methoden?	Offenes Antwortformat: Folgende Themen würde ich noch mit aufnehmen oder ausführlicher behandeln	Aktuelle Fragestellungen werden in die Veranstaltung angemessen integriert	*Nicht in Inventar*
Weitere Kommentare:	Offenes Antwortformat: Kürzungs- , Verbesserungs- und Änderungsvorschläge	Eine selbstständige und aktive Auseinandersetzung mit den Lerninhalten wird in der Veranstaltung gefördert	*Nicht in Inventar*
Welche weiteren Themen/Inhalte sollte GRADE anbieten?	Offenes Antwortformat	In der Lehrveranstaltung herrscht ein konstruktives, positives Lernklima	Die Arbeitsatmosphäre war konstruktiv

Die Aspekte des Lehrevaluationsbogens von GRADE sind alle auch in dem Inventar vorhanden. Die Frage der guten Strukturierung der Veranstaltung findet sich in dem Inventar in mehrere Teilaspekte aufgegliedert (Zeit, Stoffmenge, klare Lernziele, Organisation). Die Aussage zur „Erfüllung der Erwartungen“ wird als allgemeine Aussage zur Zufriedenheit mit der Gesamtnote gleichgesetzt.

Bezüglich des offiziellen Evaluationsbogens der Goethe-Universität haben mehrere Aspekte keine Entsprechung in dem Inventar: Die konkrete Frage, ob der Dozent auf Fragen der Teilnehmer eingeht, ob aktuelle Fragestellungen in der Veranstaltung angemessen integriert werden, und ob ein guter Überblick über das behandelte Stoffgebiet vermittelt wird. Weiterhin wird auch die Frage nach der Erkennbarkeit eines „roten Fadens“ gestellt.

8.1.6 Sind alle Inventarinhalte aus wissenschaftlicher Sicht qualitätsrelevant?

In Folge der theoretischen Ausdifferenzierung des Begriffes der Lehrqualität und der sich weiter entwickelnden Studienlage wurde die derzeitige Version des Evaluationsbogens anhand von zwei Expertinnen auf dem Gebiet der Hochschuldidaktik hinsichtlich der Angemessenheit bezüglich der Erfassung qualitätsrelevanter Aspekte begutachtet. Beide Expertinnen bescheinigten, dass die in dem Inventar enthaltenen Items alle ihre Berechtigung in einem Lehrevaluationsinventar haben. Eine Expertin merkte an, dass das Item „Ich glaube,

die heute erlernten Inhalte in meiner Dissertation umsetzen zu können“ nur zu dem spezifischen Kontext einer Promotion passe.

8.1.7 Weitere Studienergebnisse

Rindermann (2009, S. 55-56) sichtete verschiedene Studien und identifizierte drei Aspekte als besonders relevant für *gute Lehre.* Zwei dieser drei als grundsätzlich identifizierten Aspekte wurden in dem bisherigen Inventar berücksichtigt:

1. „Gute Strukturierung“ in Form der Items zu den Lernzielen, der Organisation, der Stoffmenge, des Zeitrahmens und des Unterrichtstempos.
2. „Didaktische Methodenvielfalt und –sicherheit“ in Form der Items zu den angemessenen Unterrichtsmaterialen und des „Sich Einbringenkönnens“.
3. Iteminhalte zu sozialer Kompetenz und Persönlichkeitseigenschaften wie Freundlichkeit, Offenheit und Engagement werden in dem Inventar bislang nicht abgefragt.

8.1.8 Schlussfolgerungen

Insgesamt spiegeln sich die Mehrheit der qualitätsrelevanten Aspekte des Inventars in Theorien, Befragungen und anderen Inventaren sinngemäß wider, und zwar hinsichtlich der Durchführung einer Veranstaltung (wie „Stoffmenge“ und „zeitliche Struktur“) und der Vermittlung von Lehrinhalten („klare Vermittlung von Inhalten“).

Die in dem Inventar fehlenden Aspekte wurden über die verschiedenen Evidenzquellen hinweg ihrem Inhalt entsprechend in acht Kategorien zusammengefasst (Tabelle 8).

Tabelle 8: Kategorisierung der fehlenden Inventar-Inhalte

Zusammengefasste Qualitätsdomäne	*Aspekte aus Studierendenbefragungen, Theorien und anderen Lehrevaluationsinventaren*
Das Wissens des Dozenten	• Das Wissen des Dozenten über den Unterrichtsgegenstand • Experte (Wissen über den Kursinhalt und darüber hinaus)
Soziale Kompetenzen in Form von Freundlichkeit, Respekt und Gleichbehandlung	• Freundlichkeit des Dozenten, Sorge und Respekt für Studierende • Moralisch • Auf die Studierenden zentriert
Das Erlernte umgesetzt (zum Beispiel durch Übungen) Darauf aufbauend: Rückmeldung an die Teilnehmer (eingehend)	• Übungsaufgaben • Angemessene Übung • Konsolidierung, Sicherung, intelligentes Üben • Eine selbstständige und aktive Auseinandersetzung mit den Lerninhalten wird in der Veranstaltung gefördert • Eingehend (responsive)
Roter Faden	• In der Veranstaltung sind inhaltliche Zusammenhänge („roter Faden“) deutlich erkennbar
Begeisterung oder Enthusiasmus des Dozenten	• Begeistert • Enthusiasmus des Dozenten für das Thema und für die Lehre
Erreichbarkeit und Hilfsbereitschaft des Dozenten	• Verfügbarkeit und Hilfsbereitschaft des Dozenten • Möglichkeit, auch nach Ende der Lehrveranstaltung Kontakt zum Dozenten aufnehmen zu können/Fragen klären zu können. Eventuell Hilfe bei der Umsetzung der Inhalte der Lehrveranstaltung erbitten zu können • Verbindend (connector) • Schülerorientierung
Die richtige Kursgröße	• die richtige Gruppengröße
Aktualität der Veranstaltungsinhalte	• Aktuelle Fragestellungen werden in die Veranstaltung angemessen integriert

Bei dieser Kategorisierung lag die Schwierigkeit darin, dass nicht alle genannten Aspekte vollständig identisch waren und somit Interpretationsspielraum zuließen. Folgende Aspekte werden nicht in Tabelle 8 aufgeführt, da sie mehreren Aspekten zugeordnet werden können:

1. Die Frage nach der „Erfahrung des Dozenten“: Diese wird als Teil des „Wissens des Dozenten“ angesehen, aber auch als Teil davon, verständliche Erklärungen geben zu können oder organisatorische Aspekte wie den Zeitrahmen zu berücksichtigen.
2. Auch die Aussage, ob der Veranstalter angemessen auf Fragen der Teilnehmer eingeht, wird verschiedenen Bereichen zugeordnet: Einmal als Teil des „Wissens des Dozenten“, da er die Frage inhaltlich beantworten muss und als Teil „sozialer Kompetenzen“; nämlich, ob er überhaupt und in einer angemessenen Art darauf eingeht.
3. Die „Vorbereitung des Dozenten“ wird als übergeordnete Aussage zu den Aspekten der Organisation, des Fachwissens und der Angemessenheit von Stoffmenge, Zeitrahmen sowie Unterrichtstempo angesehen.
4. Ob der Dozent Rückmeldung an die Teilnehmer gibt, wird an die Aussage zu den Übungen angeschlossen; da es sich beim Promotionskolleg um jeweils einmalige Veranstaltungen handelt, und Teilnehmer auch nur jeweils einen Kurs buchen können, gibt es pro Thema nur einen Kontakt zu dem Dozenten. Die Ergebnisse der Übungen sind somit die einzige Möglichkeit, den Teilnehmern eine fachliche Rückmeldung zu geben.

Von den in Tabelle 8 zusammengefassten Kategorien werden für die in dieser Arbeit vorgestellte spezifische studentische Lehrevaluation vier als angemessen zur Abfrage durch die Studierenden eingestuft und in einer anstehenden Revision eingefügt: Das Fachwissen des Dozenten, seine sozialen Kompetenzen, Übung inklusive Rückmeldung und der rote Faden der Veranstaltung.

Bei vier Aspekten wird auf eine Berücksichtigung in einer Revision des Inventars verzichtet: Begeisterung, Erreichbarkeit, die richtige Gruppengröße und die Aktualität der behandelten Fragestellungen.

- Ob der Dozent begeistert wirkt, wird maßgeblich aus Gründen der Sparsamkeit in Kombination mit dem Nutzen weggelassen: Es muss keine dauerhafte

Begeisterung bei den Studierenden geweckt werden, da es sich nicht um eine semesterlange Vorlesung handelt, sondern um zweistündige Veranstaltungen mit konkretem Inhalt, der praktisch angewandt werden soll.

- Die Erreichbarkeit des Dozenten wird aus ähnlichen Gründen nicht abgefragt. Zusätzliche Fragen von Teilnehmern sollen und werden über die Kontaktdaten der Promotionskollegsleitung beantwortet.
- Die richtige Gruppengröße wird auch nicht abgefragt, da es eine feste Gruppengröße gibt, die aus verschiedenen Gründen so festgelegt wurde. Falls es in Kursen zu Problemen in diesen Bereichen kommen sollte, gibt es auch ausreichend Platz in den Freitextantworten.
- Die Aktualität der Fragestellungen wird bei der Evaluation des Promotionskollegs als nicht besonders bedeutsam angesehen: Es geht mehrheitlich um die Vermittlung von Grundkenntnissen und grundlegender Kompetenzen.

8.2 Grundannahme 2: Die Items differenzieren plausibel hinsichtlich ihres Inhalts

Die zweite Grundannahme wird anhand zwei aufeinander aufbauenden Schritte überprüft: Zunächst wird anhand eines statistischen Modells geschätzt, auf welche Quellen die Varianz der einzelnen Items zurückzuführen ist. Auf Basis dieser Ergebnisse wird beurteilt, ob die einzelnen Items plausibel zwischen den verschiedenen zu evaluierenden Aspekten differenzieren: Das bedeutet einerseits, dass ein Item, das ein qualitätsrelevantes Merkmal beispielsweise hinsichtlich eines Dozenten erfassen soll, systematisch zwischen verschiedenen Dozenten Qualitätsunterschiede abbilden kann. Andererseits sollte es auch plausibel differenzieren, also nicht stärker zwischen verschiedenen Veranstaltungsthemen als zwischen Dozenten. Ansonsten müsste überdacht werden, ob das Item zum Zweck der Dozentenbeurteilung geeignet ist.

Im zweiten Schritt werden die Ergebnisse mit einer Studie verglichen, in der ebenfalls eine Varianzkomponentenschätzung an studentischen Lehrevaluationsdaten durchgeführt wurde. Hierdurch soll abgeschätzt werden, ob die Ergebnisse dieser Arbeit aus empirischer Sicht einem zu erwarteten Muster folgen.

Bislang stellte diese Grundannahme keinen elementaren Bestandteil der Validitätsbeurteilung im Kontext von Lehrevaluationsinventaren dar und wurde bisher nur in wenigen Studien berücksichtigt:

1. Rantanen (2013) errechnete einen Mittelwert basierend auf fünf Items („Erfahrung des Dozenten in dem Fachgebiet“, „Lehrfähigkeiten des Dozenten“, „visuelle Hilfestellungen“, „Interaktion mit Studierenden und Umsetzung des Erlernten“ (learning assignments) und betrachtete als Varianzquellen die Dozenten, Kurse, die Studierenden und die Interaktion aus Kursen und Dozenten.
2. Feistauer und Richter (2016) schätzten Varianzkomponenten für je vier Skalen (*Planung und Darstellung*, *Umgang mit Studierenden*, *Interessantheit und Relevanz* sowie *Schwierigkeit und Umfang*) eines Lehrevaluationsinventars: Seminare/Vorlesungen, Dozenten, Studierende und einer Interaktion aus Studierenden und Dozenten.

8.2.1 Varianzkomponentenschätzung

Eine geeignete Methode zur Überprüfung dieser zweiten Grundannahme ist die Varianzkomponentenschätzung (zum Beispiel Searle, Casella & McCulloch, 1992): Hierbei wird untersucht, ob und in welchem Ausmaß sich die Variation der Testwerte durch die Abstufungen unabhängiger Variablen erklären lassen. Mit anderen Worten: Die Summe der einzelnen Varianzkomponenten ergibt die Varianz der abhängigen Variablen.

Die Varianz von Lehrevaluationsitems soll je nach Inhalt maßgeblich auf die Teilnehmer, die Dozenten oder die Veranstaltungsthemen zurückzuführen sein, und diese sind somit als Varianzkomponenten zu spezifizieren. Interaktionseffekte zwischen Themen und Dozenten oder Dozenten und Teilnehmern sind sehr plausibel, können in diesem Design aber nicht nachgewiesen werden, da die Daten hierfür keine entsprechende Struktur aufwiesen: Nur manche Dozenten unterrichteten verschiedene Kurse, und im zeitlichen Verlauf gibt es keine Teilnehmer, die eine größere Menge der Dozenten beurteilten.

Eine Varianzkomponentenschätzung wurde für zwölf Items des Inventars durchgeführt: Elf Items mit einer vierstufigen Ratingskala sowie das sechsstufige Item

„Gesamtnote“. Items mit dreistufiger Ratingskala wurden von den Analysen aufgrund ihrer geringeren Anzahl von Abstufungen ausgenommen.

Modelle

Zur Schätzung der Varianzkomponenten wurde für jedes Item des Lehrevaluationsinventars ein *Lineares Mischmodell* spezifiziert (siehe Gleichung 8): Die Antworten auf jedes Lehrevaluationsitem repräsentieren die abhängigen Variable y_i und werden durch die Zufallsfaktoren Dozenten (d), Veranstaltungsthemen (t) und Teilnehmer (s) als unabhängigen Variablen erklärt. Das Modell enthält den Gesamtmittelwert β_0, separate Varianzkomponenten für die Zufallsfaktoren (ν_d, ν_t, ν_s) sowie ein Residuum ε_i:

$$y_{itds} = \beta_0 + \nu_d + \nu_t + \nu_s + \varepsilon_i \qquad (8)$$

Die Varianz der Antworten wird hierdurch in erklärende Varianzanteile und Residualvarianz var(ε_i) zerlegt (siehe Gleichung 9):

$$\text{var}(y) = \text{var}(\nu_d) + \text{var}(\nu_t) + \text{var}(\nu_s) + \text{var}(\varepsilon_i) \qquad (9)$$

Bei einem Mischmodell werden sowohl *Zufalls-* (die Varianzkomponenten) als auch *feste Effekte* (der Gesamtmittelwert) als unabhängige Variablen eingebracht. Zu beiden Begriffen wurden verschiedene Definitionen aufgestellt (Überblick bei Gelman, 2005). Eine geläufige Definition lautet, dass Variablen als feste Effekte zu spezifizieren sind, wenn die für diese Variable verwendete Stichprobe dem Umfang der Population entspricht. Wenn nur ein Teil der Population in die Analyse eingeschlossen wird, werden Zufallseffekte verwendet, wobei vorausgesetzt wird, dass die Stichprobe zufällig aus einer Verteilung eines Faktors gezogen wurde. Die Varianzkomponenten in dieser Analyse wurden als Zufallseffekte spezifiziert, da die Teilnehmer, Dozenten und Kursthemen langfristig betrachtet nur eine Auswahl der Gesamtmengen darstellen.

Schätzverfahren

Zur Schätzung der Varianzkomponenten wurde die *Markov Chain Monte Carlo-Methode (MCMC)* auf Basis der Bayes-Statistik genutzt.

Die Bayes-Statistik verfolgt einen deduktiv-statistischen Ansatz. Bei diesem wird die Wahrscheinlichkeit *P* einer Hypothese *H* bei gegebenen Parametern *D*

beziehungsweise Wahrscheinlichkeitsverteilungen (wie Mittelwert und Varianzen) berechnet: P(H|D). Sie stellt somit die Inversion der *Klassischen Statistik*[4] dar.

In der Bayes-Statistik können Vorwissen oder Vorannahmen integriert werden, um genauere Ergebnisse zu erhalten: Diese Annahmen werden in einer *Prior-Wahrscheinlichkeit* angegeben. Das endgültige Ergebnis nach einer bayesianischen Schätzung wird anhand einer *Posterior-Wahrscheinlichkeit* angezeigt. Falls spezifische Vorannahmen nicht integriert werden sollen, werden Verteilungsannahmen für das zu schätzende Modell als *nicht-informative Prior-Verteilung* (im Kontrast zur *informativen Prior-Verteilung* mit konkreten Annahmen) eingebracht.

MCMC-Methoden sind Sampling-Methoden, anhand derer eine Sequenz von Beobachtungen generiert wird, denen sich von einer spezifizierten multivariaten Wahrscheinlichkeitsverteilung aus angenähert wurde: Eine *Markov-Kette* leitet auf Basis einer nur begrenzt bekannten Datenlage Prognosen ab. Hierbei wird ein neuer Wert auf Basis des vorangegangenen Wertes aus der Posterior-Verteilung *gesampelt*. Dieser Prozess wird so lange betrieben bis die Lösung konvergiert. Algorithmen, nach denen diese berechnet werden, sind beispielsweise der *Metropolis-Hastings*- und ein Spezialfall von diesem der *Gibbs-Algorithmus*. Der Begriff *Monte Carlo* bezieht sich auf den Simulationsprozess.

Eine *MCMC*-Analyse wird nach Lynch (2007) in fünf Schritte unterteilt:

1. *Spezifizierung einer Likelihood-Funktion oder sampling density bei gegebenen Modellparametern:* Die Likelihood-Funktion entspricht der bedingten Wahrscheinlichkeit auf der auch *Maximum Likelihood*-Schätzungen basieren.
2. *Spezifizierung einer Prior-Verteilung für die Modellparameter:* Die Vorannahmen für die Testung der eigentlichen Hypothesen werden als Verteilungen spezifiziert.
3. *Herleitung einer Posterior-Verteilung für die Modell-Parameter* basierend auf der in den Schritten 1 und 2 angegebenen Likelihood-Funktion und Prior-Verteilung.

[4] Zur Klassischen Statistik zählen die statistischen Signifikanztests (Nickerson, 2000) und die Gruppe der *Maximum Likelihood*-Schätzalgorithmen. Bei ersterem wird die bedingte Wahrscheinlichkeit P der vorhandenen Daten D unter der Annahme der Nullhypothese H_0 berechnet: $P(D|H_0)$, und bei letzterem, unter der Annahme des Populationsparameters θ $P(D|\theta)$.

4. Simulation der Parameter, um eine Stichprobe der Posterior-Verteilung der Parameter zu erhalten
5. Zusammenfassung dieser Parameter-Samples durch deskriptive Statistiken

Wie in Abschnitt 7.4 erläutert liegen die Daten in einer nicht vollständig gekreuzten Struktur vor. Das bedeutet, dass nicht jeder Teilnehmer des Promotionskollegs an allen acht Kursen teilgenommen hat, ein Kurs nicht immer vom gleichen Dozenten gehalten wurde, und Dozenten verschiedene Kursthemen unterrichteten. Für insbesondere hierarchische Datenstrukturen ist eine Schätzung anhand der MCMC-Methode sinnvoll und hierbei robuster als nach einer Maximum Likelihood-Schätzung (Browne & Draper, 2006; Chung, Rabe-Hesketh, Dorie, Gelman & Liu, 2013).

Als Software wurde hierfür das R-Paket *MCMCglmm* (Hadfield, 2010) genutzt. Die Ergebnisse basieren auf 10000 Simulationen (Burn-in = 5000, Thinning = 10). Als nicht-informative Verteilungsannahme wurde *inverse-Wishart* gewählt. Dies wurde in diesem Fall als angemessen angesehen, da eine ausreichend große Stichprobe vorhanden ist und auch keine konkreten Vorannahmen bekannt sind. Die Syntax für ein Item findet sich als Beispiel in Anhang D.

Das Ergebnis der Schätzung besteht aus den Varianzkomponenten der unabhängigen Variablen. Für die Ergebnispräsentation kann entweder der Modal- oder Mittelwert der aus den 10000 Simulationen entstandenen Daten sein. In dieser Arbeit wird der Mittelwert mit einem entsprechenden Glaubwürdigkeitsintervall genutzt und je Komponente als prozentualer Anteil an der Gesamtvarianz dargestellt. Ein Glaubwürdigkeitsintervall kennzeichnet den Bereich eines Merkmals, in dem sich mit einer bestimmten Wahrscheinlichkeit der gesuchte Populationsparameter befindet.

8.2.2 Ergebnisse und Schlussfolgerungen

Es wurden 470 Teilnehmer, 27 Dozenten und 8 Veranstaltungsthemen eingeschlossen. Die Teilnehmer hatten 2218 Lehrevaluationsbögen ausgefüllt, von denen 38 keinen Code aufwiesen und somit aus der Analyse ausgeschlossen wurden. Im Mittel hatte ein Teilnehmer 4.6 Bögen abgegeben.

Tabelle 9: Die Ergebnisse der Varianzkomponentenschätzung (Mittelwert der Varianzkomponenten in Prozent)

Items	*Dozent*	*Kursthema*	*Teilnehmer*	*Residuum*
Die Lernziele der Veranstaltung waren klar erkennbar	14.474	6.337	8.713	70.477
Ich hatte die Möglichkeit, mich aktiv an der Veranstaltung zu beteiligen	32.825	0.025	9.507	57.643
Die Arbeitsatmosphäre war konstruktiv	24.416	0.046	12.811	62.728
Die Erklärungen des Dozenten waren verständlich	21.481	10.263	11.412	56.845
Die Durchführung durch den Dozenten war motivierend	24.785	0.053	11.58	63.582
Die verwendeten Unterrichtsmaterialien waren angemessen	14.094	1.105	14.302	70.499
Ich hatte ausreichende Vorkenntnisse, um der Veranstaltung zu folgen	6.538	6.822	26.641	60
Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können	9.472	16.331	13.902	60.295
Die Inhalte der heutigen Veranstaltung sind für mich persönlich relevant	8.45	17.182	15.197	59.171
Die Veranstaltung war gut organisiert	16.367	0.246	10.289	73.099
Ich würde diese Veranstaltung anderen Doktoranden weiterempfehlen	22.359	1.201	11.287	65.153
Gesamtnote für diese Veranstaltung (in Schulnoten)	28.369	0.034	13.327	58.27

Für alle zwölf untersuchten Items konnte nachgewiesen werden, dass sie differenzieren (Tabelle 9). Auch der Nachweis einer plausiblen Differenzierung ist gegeben. Dies wird in Abbildung 15 veranschaulicht, in dem die einzelnen Komponenten ohne Residuen dargestellt und jeweils am Gesamtanteil erklärter Varianz relativiert werden: Zum Beispiel geht ein Großteil der erklärten Variation der Antworten zu den *Vorkenntnissen* auf die unterschiedlichen Teilnehmer selbst zurück. Das spricht dafür, dass das Item in der Lage ist, diese zu erfassen. Ebenso plausibel ist es, dass die beiden höchsten Anteile erklärter Varianz bei der Aussage zur *Umsetzbarkeit* auf das Thema und die Teilnehmer rückführbar sind. So ist es in der Lage auszusagen, dass einige Themen allgemein besser umgesetzt werden können, und gleichzeitig manche Teilnehmer alle Kurse besser oder weniger gut für Ihre Arbeit nutzen können. Ebenso ist eine großteilige Differenzierung nach dem Dozenten

bei dem Item bezüglich der *Beteiligung* plausibel. Als eher unplausibel würde eine großteilige Variation nach den Teilnehmern erscheinen. Über alle Items hinweg zeigt sich ein hoher Anteil an Residualvarianz. Dieser besteht wahrscheinlich großteilig aus Interaktionen der drei unabhängigen Variablen. Die Ergebnisse der Varianzkomponentenschätzung werden detailliert in Anhang E dargestellt.

Als Einschränkung dieser Ergebnisse kann zum einen gesehen werden, dass aufgrund der Datenlage Interaktionseffekte nicht geschätzt werden konnten. In dieser Situation wären diese nur mit einem Paralleltestdesign zu identifizieren, was mit einem hohen Aufwand verbunden ist (Aufwand für die Konstruktion des Paralleltest und Teilnehmer brauchen mehr Zeit zum Ausfüllen). Als weitere Einschränkung kann angesehen werden, dass das Schätzverfahren auf ein metrisches Skalenniveau der abhängigen Variablen ausgerichtet ist. Die in der Analyse eingeschlossenen Items besaßen allerdings ein vierstufiges Ordinalskalenniveau und das der Gesamtnote ein sechsstufiges. Daher wurden die drei dreistufigen Variablen vorab ausgeschlossen.

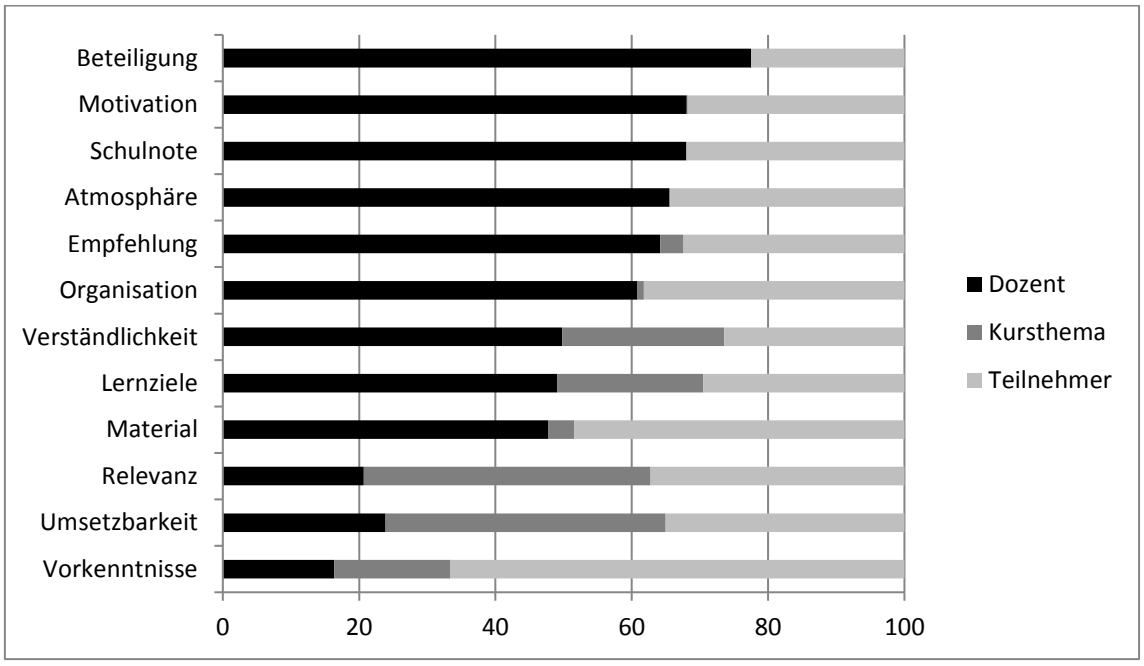

Abbildung 15: Varianzkomponenten in Prozent ohne Residuen (Anteil erklärter Varianz = 100%)

8.2.3 Vergleich mit anderen Studienergebnissen

Als Ergänzung zur Varianzkomponentenschätzung werden deren Ergebnisse mit denen vorangegangener Studien mit ähnlicher Methodik verglichen. Dadurch kann

abgeschätzt werden, ob beispielsweise der hohe Anteil an Residualvarianz typisch für studentische Lehrevaluationsergebnisse ist.

Die Studie von Feistauer und Richter (2016) untersuchte die Inter-Rater-Reliabilität der Lehrqualität anhand studentischer Evaluationen. Die Inter-Rater-Reliabilität wurde als der Anteil der Gesamtvarianz studentischer Evaluationen definiert, die durch „Kurse“ und „Dozenten“ erklärt wird. Diese beiden Varianzkomponenten wurden als *Target Variance* gesehen, die den Grad angeben, nach dem verschiedene Studierende diese *Targets* in gleicher Weise beurteilen. Die Reliabilität würde durch die Steigerung des Anteils der *Target Variance* erhöht.

Hierfür wurden die Beurteilungen der Kurse durch Studierende hinsichtlich des Konstrukts *Effektivität der Lehre (teaching effectiveness)* genutzt. Die Daten wurden anhand der Inventare FEVOR (Fragebogen zur Evaluation von Vorlesungen) und FESEM (Fragebogen zur Evaluation von Seminaren) mit je vier Skalen und zwei globalen Ratings (je eines für Kurse und eines für Dozenten) erhoben (Staufenbiel, 2000).

Als Varianzkomponenten wurden Dozenten, Kurse und Studierende als Zufallsfaktoren spezifiziert. In einem weiteren Modell wurde die Interaktion aus Studierenden und Dozenten mit aufgenommen. Kurse wurden nach Vorlesungen und Seminaren getrennt analysiert. Als Ergebnis zeigte sich, dass die Dozenten und Kurse essenzielle Varianzquellen für alle vier Facetten und die globalen Urteile sind.

Da die Kurse des Promotionskollegs eher einem Seminar als einer Vorlesung ähneln, wurden deren Ergebnisse für den Vergleich herangezogen (siehe Tabelle 10). Die vier Skalen und zwei globalen Urteile des FESEM wurden jeweils mit einem inhaltlich relativ ähnlichen Einzelitem des Promotionskolleg-Inventars verglichen. Allerdings können die Ergebnisse dieser Arbeit mit denen der Studie nicht direkt verglichen werden: Die Veranstaltungen unterscheiden sich (Einzel-Kurse versus Seminare), und die Inhalte sind unterschiedlich (Einzelitems versus aggregierte Items in Skalen). Hierbei kommt noch hinzu, dass innerhalb der Skalen manche Items relativ heterogen sind und nicht unbedingt miteinander korrelieren müssen.

Dennoch können gewisse Gemeinsamkeiten ausgemacht werden, wie der hohe Anteil des Residuums. In Tabelle 10 werden die Varianzkomponenten des Modells

mit Interaktionseffekt dargestellt und in der letzten Zeile die Höhe des Residuums aus dem Modell ohne Interaktion. Die Varianzkomponenten beider Modelle mit und ohne Interaktion sind sich sehr ähnlich; die Varianzkomponente der Interaktion speist sich maßgeblich aus dem Residuum.

Der Vergleich im Einzelnen:

1. Die erste Skala beinhaltet unter anderem Items zu klarer Gliederung, Klarheit und Verständlichkeit des Dozenten oder Hilfsmittel in guter Qualität und wird mit dem Item zur Verständlichkeit des Dozenten verglichen. Dieses hat wie zu erwarten eine hohe auf die Dozenten rückführbare Varianz. Die Varianz der Skala lässt sich hauptsächlich auf Studierende oder deren Interaktion mit den Dozenten zurückführen.
2. Die zweite Skala fragt unter anderem Aspekte zur Freundlichkeit des Dozenten ab, und ob er auf Fragen der Studierenden eingeht. Diese wird mit der Aussage zur konstruktiven Atmosphäre verglichen. Bei dem Item des Promotionskollegs geht die Varianz am stärksten auf die Dozenten zurück, während in der Skala alle drei Komponenten eher ähnlich sind; nur die Interaktion ist etwas höher ausgeprägt.
3. Drei der Varianzkomponenten sind ähnlich, aber bezüglich der Dozenten unterscheiden sie sich deutlich: In der Skala werden maßgeblich dozentenbezogene Aspekte abgefragt, was den entsprechenden Varianzanteil erklärt („Der Dozent gestaltet die Veranstaltung interessant, und das Seminar ist für die spätere Berufspraxis nützlich“). In dem Item aus dem Inventar des Promotionskollegs ist die Varianz maßgeblich auf das Veranstaltungsthema und die Studierenden selbst zurückzuführen
4. Diese Skala beinhaltet Aspekte zu Schwierigkeit, Stoffumfang und Tempo. Hier wird das Item zu den ausreichenden Vorkenntnissen als Vergleich herangezogen. In diesem Item wird Varianz maßgeblich durch die Studierenden erklärt, was plausibel ist. Stoffumfang und Tempo sind aber auch von den Dozenten anhängig, was diese Komponenten in der Skala ausgeprägter werden lässt. Diese beiden Aspekte des Inventars des Promotionskollegs wurden in dieser Arbeit nicht einer Varianzkomponentenschätzung unterzogen.

5. Bei der Gesamtnote der Promotionskollegsbeurteilung kommt die Varianz maßgeblich durch die Dozenten zustande. Dies kann darauf zurückzuführen sein, dass eine starke Fokussierung in den einmalig zweistündigen Kursen auf den Dozenten besteht. Der etwas schwächere Anteil der Dozenten bei der globalen Beurteilung bei Feistauer et al. kann bedeuten, dass in einem Seminar verschiedene Faktoren die Zufriedenheit beeinflussen.

Zusammengefasst kann ausgesagt werden, dass heterogene Skalen nicht direkt mit Einzelitems vergleichbar sind. Zumindest kann dargestellt werden, dass ein hoher Anteil an Residualvarianz als üblich anzusehen ist, und dass dieser durch die Spezifikation von Interaktionseffekten verringert werden kann.

Als Ergänzung kann noch die Studie von Rantanen (2013) hinzugezogen werden, bei dem fünf Lehrevaluations-Items aggregiert wurden: Auch hier zeigte sich mit 46.4% ein hohes Residuum, Dozenten hatten einen Anteil von 24.6%, die Studierenden 16.8%, die Kurse 6% und eine Interaktion aus Kursen und Dozenten 5.3%.

Tabelle 10: Vergleich der Varianzkomponenten (in Prozent) der FESEM-Skalen mit jeweils einem Item des in dieser Arbeit untersuchten Inventars des Promotionskollegs (PK).

	Planung und Darstellung (FESEM)	*Die Erklärungen des Dozenten waren verständlich (PK)*	*Umgang mit Studierenden (FESEM)*	*Die Arbeits-atmosphäre war konstruktiv (PK)*	*Interessant-heit und Relevanz (FESEM)*	*Die Inhalte der heutigen Veranstalt-ung sind für mich persönlich relevant (PK)*	*Schwierigkeit und Umfang (FESEM)*	*Ich hatte ausreichende Vorkennt-nisse, um der Veranstalt-ung zu folgen (PK)*	*Schulnote für Kurs (FESEM)*	*Schulnote für Dozent (FESEM)*	*Schulnote (PK)*
Studierende/Teilnehmer	21.1	11.412	16.5	12.811	16.2	15.197	13.9	26.641	10.7	10.6	13.327
Dozenten	7.1	21.481	14.8	24.416	16.5	8.45	17.3	6.538	10.5	17.5	28.369
Kurse/Kursthema	13.1	10.263	18.6	0.046	15	17.182	8	6.822	13.7	14.9	0.034
Interaktion Studierende/Dozent	20.1	-	26.3	-	18	-	7.5	-	16.9	24.2	-
Residuum	38.6	-	23.7	-	34.4	-	53.4	-	48.2	32.7	-
Residuum (aus Modell ohne Interaktion)	58.1	56.845	49.3	62.728	51.8	59.171	60.5	60	64.7	56.2	58.27

8.3 Validitäts-Argumentation

Unter Validitäts-Argumentation versteht Kane (2013) die Evaluation sämtlicher Ansprüche: Die vorgeschlagenen Interpretationen und Verwendungen seien in dem Ausmaß valide, wie die Argumentation für die Interpretationen und Verwendungen vollständig, klar und kohärent ist sowie ihre Schlüsse angemessen und die Annahmen, die die Rechtfertigungen dieser Schlüsse unterstützen, aus sich heraus plausibel sind oder durch Evidenzen gestützt werden.

Laut den *Standards* (AERA et al., 2014) integriert eine gut fundierte Validitäts-Argumentation verschiedene Stränge von Evidenz in eine kohärente Darstellung des Ausmaßes, in dem Evidenz und Theorie die beabsichtigten Interpretationen von Testwerten für eine bestimmte Verwendung unterstützen.

Die beabsichtigte Interpretation der Ergebnisse des in dieser Arbeit betrachteten Lehrevaluationsinventars lautet, dass höhere Ergebniswerte ein höheres Ausmaß der qualitätsbezogenen Zufriedenheit von Teilnehmern mit der Durchführung eines Unterstützungsprogramms für Promovierende und mit der entsprechenden Vermittlung von Lehrinhalten darstellen. Im Sinne dieser Interpretation wurden zwei Grundannahmen aufgestellt und anhand verschiedener Evidenzquellen überprüft: Zunächst, dass wenn die Ergebnisse als Zufriedenheit von Teilnehmern einer Lehrveranstaltung angesehen werden, diese Ergebnisse alle qualitätsrelevanten Zufriedenheitsaspekte berücksichtigen. Weiterhin sollen die entsprechenden Items des Inventars die beabsichtigten Unterschiede hinsichtlich der qualitätsbezogenen Zufriedenheit abbilden können.

8.3.1 Grundannahme 1

Zunächst wurden die Iteminhalte des Inventars in seiner bisherigen Form mit Ergebnissen von Studien verglichen, in denen Studierende selbst angaben, was sie unter guter Lehre verstehen. Da sich diese Angaben auf allgemeine Lehrveranstaltungen beziehen, besteht die Möglichkeit, dass für ein Unterstützungsprogramm von Promovierenden auch andere und/oder zusätzliche Aspekte relevant sind. Dementsprechend wurde eine Umfrage unter den Teilnehmern des Promotionskollegs gestartet, in der sie die bisherigen Iteminhalte hinsichtlich ihrer subjektiven Relevanz bewerten und mit von ihnen selbst als relevant eingestuften Aspekten ergänzen sollten.

Anhand dieser beiden Evidenzquellen wurde die Perspektive der Teilnehmer erfragt. Allerdings ist davon auszugehen, dass Teilnehmer von Lehrveranstaltungen nicht über einen allgemeinen Überblick hinsichtlich didaktischer Methoden verfügen. Dementsprechend wurden als Ergänzung weitere Evidenzquellen herangezogen: Einerseits wurden die bisher bestehenden Inhalte des Inventars mit denen in Lehr-Lern-Theorien aufgeführten relevanten Aspekten und mit einer Zusammenfassung verschiedener Studienergebnisse verglichen. Darüber hinaus wurden die Iteminhalte des bisherigen Inventars von zwei Expertinnen der Hochschuldidaktik hinsichtlich ihrer Relevanz als Maß von Lehrqualität eingeschätzt. Ebenso wurden zum Vergleich zwei andere Lehrevaluationsinventare hinzugezogen, bei dem eines sich ebenfalls mit einem Lehrangebot für Promovierende befasst.

Zusammengefasst hat sich unter Berücksichtigung dieser Evidenzquellen bestätigt, dass die bisherigen Iteminhalte des Inventars relevante Maße von Lehrqualität aus Sicht von Lehr-Lern-Theorien, Experten und Teilnehmern sind. Darüber hinaus zeigte sich, dass acht Aspekte in dem Inventar nicht berücksichtigt, aber jeweils mindestens in einer Evidenzquelle als qualitätsrelevant bezeichnet wurden. Vier dieser Aspekte werden bei einer anstehenden Revision des Inventars eingearbeitet. Bei vier Aspekten wird von einer Übernahme abgesehen, da diese in dem spezifischen Kontext des Promotionskollegs als nicht angemessen betrachtet werden.

Es ist allerdings anzumerken, dass die verschiedenen Evidenzquellen nicht konkret die *Zufriedenheit* von Veranstaltungsteilnehmern darstellen, sondern auf verschiedenen Fragestellungen basierten (effektive Dozenten, Lernerfolg oder grundsätzlich auf schulischen Unterricht bezogen). Allerdings sind diese Zusammenstellungen eine wichtige Hilfestellung, um einen fundierten Überblick zu relevanten Aspekten von Lehrqualität zu erhalten. Ob die jeweiligen Aspekte auch für die eigene zu evaluierende Veranstaltung gelten, muss dann jeweils überprüft werden (wie durch eine Teilnehmerbefragung oder eine theoretische Diskussion).

Zusammengefasst wird der Anspruch bestätigt, dass die Inhalte des Inventars für die Zufriedenheit von Teilnehmern eines Unterstützungsprogramms für Promovierende relevant sind. Weiterhin hat sich gezeigt, dass im Sinne einer *Construct Deficiency* einzelne Aspekte bislang nicht in dem Inventar abgefragt wurden, und eine Überarbeitung notwendig ist.

Als weitere Evidenzquelle ist angedacht, Teilnehmer des Promotionskollegs anhand von Interviews über ihren Bedarf hinsichtlich der Vermittlung von Lehrinhalten und der Durchführung eines Unterstützungsprogramms für Promovierende zu befragen.

8.3.2 Grundannahme 2

In der ersten Grundannahme wurden die bisherigen Iteminhalte der bisherigen Fassung als relevant für die Teilnehmerzufriedenheit bestätigt. In der zweiten Grundannahme wurden die Inhalte in ihrer Form als konstruierte Items hinsichtlich psychometrischer Eigenschaften überprüft. Da sie Unterschiede hinsichtlich der qualitätsbezogenen Zufriedenheit abbilden sollen, muss ihre Varianz plausibel auf den jeweilig interessierenden Inhalt zurückzuführen sein (beispielsweise, dass ein dozentenbezogenes Item maßgeblich zwischen verschiedenen Dozenten Unterschiede abbilden kann).

Hinsichtlich dieser Annahme wurden zwei aufeinander aufbauende Evidenzquellen herangezogen: Anhand der bislang erhobenen Daten wurden die Items jeweils einzeln einer Varianzkomponentenschätzung unterzogen, um zu identifizieren, in welchem Ausmaß die Komponenten einer Lehrveranstaltung ihre Varianz bedingt. Die Ergebnisse wurden zunächst per Augenschein hinsichtlich ihrer Plausibilität bewertet. Darauf aufbauend wurden diese Ergebnisse mit den Varianzkomponentenschätzungen einer anderen Studie verglichen. Damit sollte überprüft werden, ob bestimmte Variationsmuster als typisch für Lehrqualitäts-Beurteilungen von Lehrveranstaltungsteilnehmern angesehen werden können.

Als Ergebnis kann festgehalten werden, dass vom Augenschein her die Items plausibel hinsichtlich Dozenten, Veranstaltungsthemen oder den Teilnehmern variieren. Der Vergleich mit dem Studienergebnis von Feistauer und Richter (2016) zeigte nur bedingt eine Übereinstimmung: Der hohe Anteil an Residualvarianz ist sehr ähnlich, während sich andere Komponenten unterschieden.

Als Schwächen der empirischen Prüfung der zweiten Grundannahme kann einmal der hohe Anteil nicht erklärter Varianz genannt werden (Range von 56.845% bis 73.099%). Diese ist wahrscheinlich auf die nicht erfolgte Spezifikation der Interaktionseffekte zurückzuführen. Diese wären nur durch Konstruktion eines Paralleltests möglich gewesen, um systematische von unsystematischer Varianz zu trennen. Wie an der Studie von Feistauer und Richter (2016) zu sehen ist, ist die

Annahme eines Studierenden-Dozenten-Interaktionseffekts plausibel. Für das Promotionskolleg kann zusätzlich auch ein Teilnehmer-Kurs-Interaktionseffekt angenommen werden, da manche Themen für bestimmte Dissertationen bedeutsamer sind oder ein unterschiedliches Vorwissen besteht - was beispielsweise den Umgang mit Textformatierungs- oder Datenmanagementsprogrammen betrifft. Ebenfalls sind auch dreifach Interaktionen möglich: Komplexere Themengebiete (wie die Kurse zur Klinischen Epidemiologie) können für manche Studierende bei bestimmten Lehrstilen eines Dozenten noch schwerer verständlich sein.

Eine weitere Schwäche der Analyse besteht in ihrer Ausrichtung auf intervallskalierte Daten hinsichtlich der abhängigen Variable: Die analysierten Items wiesen dagegen ein vierstufiges Ordinalskalenniveau auf und bei der Gesamtnote ein sechsstufiges. Dementsprechend wurden die drei dreistufen Items nicht in dieser Analyse berücksichtigt.

Grundsätzlich ist allerdings ein direkter Vergleich mit der Studie von Feistauer und Richter (2016) nur bedingt möglich, denn das Studiendesign unterscheidet sich hinsichtlich verschiedener Aspekte: Es wurden Analysen auf Skalen- statt auf Einzelebene durchgeführt. In den Skalen wurden auch teilweise heterogene Aspekte abgefragt. Zusätzlich handelte es sich um reguläre Seminare und nicht um ein spezifisches anwendungsbezogenes Programm für Promovierende. Damit waren zwar wegen der Möglichkeit zur studiumsbegleitenden Promotion mehrheitlich, aber nicht nur Studierende in die Analyse dieser Arbeit eingegangen (genaue Anzahl wurde nicht erfasst, da nicht abgefragt).

Für eine weitere Überprüfung der zweiten Grundannahme sind in Zukunft mehrere Aspekte geplant: Zur Varianzkomponentenschätzung soll ein Verfahren herangezogen werden, dass ordinalskalierte Daten berücksichtigen kann. Damit können auch die dreistufigen Items in die Analyse einbezogen werden. Ob ein Paralleltest konstruiert und zumindest zweitweise eingesetzt wird, muss noch hinsichtlich der Praktikabilität diskutiert werden. Weitere Vergleiche mit anderen Studien zur Schätzung von Varianzkomponenten sind abzuwarten, da anscheinend noch keine weiteren auf Einzelitemebene durchgeführt wurden.

Die auf Basis der ersten Grundannahme revidierte Fassung ist dann ebenfalls in der beschriebenen Weise einer Analyse zu unterziehen.

8.3.3 Weitere Grundannahmen im Sinne der Interpretation

Neben den beiden aufgeführten Grundannahmen hinsichtlich der Validität der Interpretation kann eine dritte getestet werden: Wenn die kausal-formativen Items des Inventars das Ausmaß der Lehrveranstaltungszufriedenheit bestimmen, dann hängen diese statistisch mit den reflektiven Maßen des Konstruktes zusammen (wie der Gesamtnote).

Diese Grundannahme bedeutet, dass beispielsweise die Aspekte der Verständlichkeit des Dozenten und der Organisation eines Kurses das Ausmaß der Zufriedenheit bestimmen und somit auch statistisch mit ihm zusammenhängen sollten. Das Ausmaß der Zufriedenheit wird in Form globaler Aussagen wie der Gesamtnote oder einer Empfehlung der Veranstaltung gemessen. Diese Grundannahme kann anhand einer Regressionsanalyse überprüft werden.

8.3.4 Die Validität der Verwendung der Ergebnisse

Die drei bislang erwähnten Grundannahmen bezogen sich auf die Interpretation der Ergebnisse: Wurden alle Zufriedenheitsaspekte abgefragt, bilden die entsprechenden Items Zufriedenheitsunterschiede plausibel ab, und hängen diese mit einer allgemeinen Zufriedenheitsaussage zusammen?

Allerdings sollen Lehrevaluationsergebnisse auch für bestimmte Zwecke verwendet werden: Hier ist insbesondere die Modifikation von Lehrveranstaltungen zu nennen. Das bedeutet zum Beispiel, dass Unterrichtsmaterialen angemessener gestaltet, die Lernziele klarer benannt werden oder die Vorkenntnisse der Teilnehmer zu berücksichtigen sind. Dementsprechend ist auch die Validität der Verwendung von Lehrevaluationsergebnissen relevant: Eine maßgebliche Grundannahme für die Verwendung lautet, dass wenn Lehrevaluationsergebnisse als eine Grundlage für Entscheidungen zur Modifikation von Veranstaltungen verwendet werden, diese mit den Beurteilungen von Didaktik-Experten übereinstimmen sollten. Dies kann in folgender Weise überprüft werden: Die Experten beurteilen die Veranstaltungen anhand derselben Kriterien wie die Teilnehmer. Im Anschluss wird die Übereinstimmung der Beurteilungen von Experten und denen der Teilnehmer berechnet.

8.3.5 Die Validität der beabsichtigten Konsequenzen

Neben der Interpretation und der Verwendung haben Lehrevaluationsergebnisse beabsichtigte Konsequenzen, die ebenso zu validieren sind, aber in dieser Arbeit nicht

berücksichtigt wurden: Auf Basis einer gegebenen Interpretation werden die Testwerte für Entscheidungen verwendet. Diese Entscheidungen beinhalten die Frage, ob und welche Aspekte einer Lehrveranstaltung verändert werden. Diese Entscheidungen haben Konsequenzen, da Lehrveranstaltungen mindestens einen Zweck zu erfüllen haben: Im Falle des Promotionskollegs lautet dieser, den Promotionsprozess für die Promovierenden selbst und die Betreuer zu erleichtern, sowie damit im besten Fall die Qualität der Dissertation zu erhöhen. Daraus lassen sich drei zu überprüfende Grundannahmen ableiten: Wenn bessere studentische Lehrevaluationsergebnisse des Promotionskollegs einen in verschiedener Hinsicht positiveren Verlauf des Promotionsprozesses zur Konsequenz haben, dann sollten diese Ergebnisse mit einer entsprechend höheren Zufriedenheit der Promovierenden und Betreuer mit dem Promotionsprozess sowie mit einer höheren Dissertationsqualität einhergehen.

Zur Überprüfung dieser Grundannahmen müssten hierfür höhere Lehrevaluationsergebnisse mit niedrigeren hinsichtlich verschiedener Aspekte verglichen werden: Fiel den Teilnehmern, die einen Datenmanagement-Kurs besser evaluierten, der Umgang mit Daten leichter als denen, die einen Kurs schlechter evaluierten? Dabei wären verschiedene Probleme zu klären: Werden hierfür nur die Gesamtnote oder einzelne Kriterien verwendet, und wie werden entsprechende Merkmale im Promotionsprozess selbst gemessen? Reicht hierfür eine globale Aussage aus, oder sollten Gespräche in Form von Interviews geführt werden? Weiterhin müssten noch weitere Variablen neben den Lehrevaluationsergebnissen berücksichtigt werden: Wie zufrieden waren die Promovierenden mit den Betreuern, wie schwierig war das Promotionsthema, und gab es weitere Probleme im Promotionsprozess?

8.3.6 Schlussfolgerung

In dieser ersten Beurteilung kann geschlussfolgert werden, dass die Validität hinsichtlich der beabsichtigten Interpretation weitgehend, aber noch nicht vollständig bestätigt werden kann: Alle bislang in dem Inventar aufgeführten Aspekte sind im Sinne der beabsichtigten Interpretation zu betrachten. Die diesen Aspekten entsprechenden Items variieren auch angemessen hinsichtlich der Qualitätsunterschiede, die sie erfassen sollen. Vier qualitätsbezogene Aspekte wurden in dem Inventar als fehlend identifiziert, werden in einer revidierten Fassung berücksichtigt und anschließend ebenfalls hinsichtlich der zweiten Grundannahme nachträglich überprüft. Ebenfalls steht auch

eine Überprüfung der Frage aus, ob die Inhalte des Inventars statistisch mit einer Zufriedenheitsbeurteilung zusammenhängen.

Diese Arbeit bezog sich auf den Nachweis von Validität hinsichtlich der Interpretation. Wenn nach dem eben skizzierten weiteren Verlauf die Validitätsbeurteilung weiterhin positiv ausfällt, sind Nachweise der Validität hinsichtlich der Verwendung und der Konsequenzen erforderlich. Entsprechende Grundannahmen wurden in diesem Kapitel vorgestellt und sind in einer auf dieser aufbauenden Arbeit zu überprüfen. Eine Übersicht dieser Grundannahmen wird in Tabelle 11 dargestellt.

Tabelle 11: Argumentationsstruktur für noch ausstehende Grundannahmen sowie die Entsprechung ihrer Evidenzen nach den Standards

Anspruch	Wenn die Aspekte in dem Inventar relevant für die Zufriedenheit mit einer Lehrveranstaltung sind, …	Wenn Lehrevaluationsergebnisse zu Entscheidungen zur Modifikation der Lehrveranstaltung verwendet werden, ….	Wenn studentische Lehrevaluationsergebnisse des Promotionskollegs besser ausfallen, …		
Daten	… dann sollten sie statistisch mit einer allgemeinen Aussage zur Zufriedenheit zusammenhängen	… dann sollten sie mit Beurteilungen von Experten hinsichtlich der Qualitätsbeurteilung übereinstimmen	… dann ergibt sich als Konsequenz eine formal bessere Dissertationsqualität.	… dann ergibt sich als Konsequenz für den Promovierenden eine höhere Zufriedenheit mit dem Promotionsprozess.	… dann ergibt sich als Konsequenz eine höhere Zufriedenheit des Betreuers mit dem Promotionsprozess.
Rechtfertigung	Regressionsanalyse	Korrelation	Zusammenhangsanalyse		
Unterstützung	Spezifikation der kausal-formativen Items als unabhängige Variable und die Gesamtbeurteilung als abhängige Variable	Abgleich von LE-Ergebnissen pro einzelnem Kurs mit Urteil eines Experten	Vergleich hoher, niedriger LE-Ergebnisse mit Qualität der Dissertation	Vergleich der Lehrevaluationsergebnisse mit Angaben des Promovierenden zu seiner Zufriedenheit.	Vergleich der Lehrevaluationsergebnisse mit Angaben des Betreuers zu Zufriedenheit.
Qualifizierer	Statistisch signifikant	-	-	-	-
Widerlegung	-	-	Unter anderem Schwierigkeit des Promotionsthemas, Qualität der Betreuung		
Daten im Sinne der Widerlegung	-	-	Befragungen von Promovierenden, Betreuern und Experten des jeweiligen Fachgebiets		
Evidenzquelle nach Standards	"Evidence based on internal structure"	"Evidence regarding relations to other variables"	"Evidence based on validity and consequences of testing"		

9. Allgemeine Diskussion und Schlussfolgerungen

9.1. Beantwortung der Fragestellung

Die Fragestellung dieser Arbeit lautet, welche Art von Validierungsprozess für studentische Lehrevaluationsergebnisse geeignet ist. Zu ihrer Beantwortung wurden drei Aspekte beleuchtet: Was soll durch die Evaluation gemessen werden? Wie wurden bisherige Validierungsstudien durchgeführt? Sind diese angemessen, und welche Alternativen gibt es?

Für die Beantwortung des ersten Aspekts zeigte sich, dass sich das durch Lehrevaluationen erfasste Merkmal je nach Evaluationsziel unterscheiden kann, und von den Institutionen konkret benannt und transparent beschrieben werden soll. Das in dieser Arbeit präsentierte Lehrevaluationsinventar wird zur Erfassung der Zufriedenheit von Teilnehmern eines Unterstützungsprogramms mit der Vermittlung von Lerninhalten und Durchführung von Lehrveranstaltungen eingesetzt. Zu ihrer Erfassung werden verschiedene qualitätsrelevante Kriterien abgefragt.

Hinsichtlich des zweiten und dritten Aspekts kann festgehalten werden, dass bisherige Validierungsstudien zu Lehrevaluationsinventaren maßgeblich von der Suche nach beziehungsweise der Überprüfung einer angenommenen Faktorstruktur sowie einer Messung des Messfehlers geprägt sind. Allerdings sind diese Methoden häufig aufgrund der meist als formativ zu bezeichnenden Item-Struktur von Lehrevaluationsinventaren unangemessen.

Denn der Einsatz dieser Methoden kann beispielsweise dazu führen, dass Items entfernt werden, da sie auf keinen Faktor laden, oder dass Skalen anhand unter Umständen redundanten Items verlängert werden, um die interne Konsistenz zu erhöhen. Erschwerend kommt hinzu, dass die mangelnde Berücksichtigung der Trennung des zu messenden Konstrukts (auf Ebene der Lehreinheit) und der Messung (auf Ebene der Studierenden) zu fehlerhaften Ergebnissen der Faktorenanalyse führen kann. Ein grundlegendes Problem ist die weit verbreitete Annahme, Validierungsstudien bestünden aus Elementen, wie sie in der Persönlichkeitspsychologie angewendet werden. Weitere Validitätsüberprüfungen suchten den Vergleich anhand eines Kriteriums, wie Lernerfolg oder der Beurteilung durch andere Personengruppen. Doch

es zeigte sich, dass es für das in dieser Arbeit erfasste Konstrukt kein angemessenes Kriterium gibt.

Als Alternative konnten argumentationsbasierte Validitätsansätze angeführt werden, die eine differenzierte und spezifische Validierung ermöglichen. Dabei stellen sie eine angemessene Alternative zu dem Konstruktmodell dar, da differenzierte Theorien nicht vorhanden oder nötig sind. Somit können sie anwendungsbezogen auf das jeweilige Instrument ausgerichtet werden.

9.2. Das Lehrevaluationsinventar in dieser Arbeit

Diese Arbeit bezieht sich hinsichtlich der theoretischen und praktischen Reflektion auf ein vorab konstruiertes und bereits in der Anwendung befindliches Lehrevaluationsinventar. Daher kann kein vollständiger Validierungsprozess ab der Intention, einen Test anzuwenden, beschrieben werden, sondern es erfolgte eine Überprüfung nach mehreren Jahren der Anwendung. Dies würde bei einer weiteren Verbreitung des argumentationsbasierten Validitätsansatzes nicht unüblich sein, da – weit über das Spektrum studentischer Lehrevaluationen hinaus – in sehr vielen Bereichen etablierte und gegebenenfalls auch standardisierte Inventare bereits eingesetzt werden. Falls der Bedarf für eine erneute Validitätsüberprüfung nach dem argumentationsbasierten Ansatz angemeldet würde, fänden daher viele Studien an schon entwickelten Inventaren statt.

Aufgrund des Umfangs solch einer Validierung auf verschiedenen Ebenen (Interpretation, Verwendung und Konsequenzen) und der entsprechenden empirischen Überprüfung konnte diese hier nicht abschließend dargestellt werden. Dies zeigt ein Problem argumentationsbasierter Ansätze auf: Konkrete Annahmen hinsichtlich der Interpretation, Verwendung und Konsequenzen können vielfältig und ihre Überprüfung mit erheblichem Aufwand verbunden sein.

9.3. Schlussfolgerung

Abschließend kann festgestellt werden, dass anhand der Orientierung an argumentationsbasierten Ansätzen Defizite vorangegangener „klassischer“ Validierungsprozesse in theoretischer wie praktischer Hinsicht aufgehoben, und eine pragmatische Vorgehensweise entwickelt werden kann. Dies könnte die Entwicklung und Anwendung studentischer Lehrevaluationsinventare positiv beeinflussen.

Anwender und Entwickler sollten demensprechend bei der Auswahl schon bestehender Inventare oder der Konstruktion von Inventaren auf einen fundierten Validierungsprozess achten, da Lehrevaluationsergebnisse verschiedene bedeutsame Folgen haben können. Dieser Validierungsprozess sollte eine stringente und evidenzgestützte Argumentation der Evaluationsergebnisse von deren Interpretation bis zu den auf ihnen basierenden Entscheidungen aufweisen.

Literaturverzeichnis

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30* (2), 221–227.

Abrami, P. C., d'Apollonia, S. & Cohen, P. A. (1990). Validity of Student Ratings of Instruction: What we know and what we do not. *Journal of Educational Psychology, 82* (2), 219–231.

Abrami, P. C., d'Apollonia, S. & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Hrsg.), *Higher Education: Handbook of Theory and Research* (Bd. 11, S. 213–264). New York: Agathon Press.

Abrami, P. C., d'Apollonia, S. & Rosenfield, S. (1997). The Dimensionality of Student Ratings of Instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (321-367). New York: Agathon Press.

Abs, H. J., Merki, K. & Klieme, E. (2006). Grundlegende Gütekriterien für Schulevaluation. In W. Böttcher (Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele* (Grundlagentexte Pädagogik, S. 97–108). Weinheim: Juventa-Verlag.

Altbach, P. G. & Selvaratnam, V. (Eds.). (1989). *From dependence to autonomy. The development of Asian universities*. Dordrecht: Kluwer Academic Publishers.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

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

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: AERA Publications Sales.

American Psychological Association. (1974). *Standards for Educational and Psychological Tests. Washington*, D.C.: American Psychological Association.

Anastasi, A. (1984). *Psychological Testing* (5. Aufl.). New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37* (1), 1–16.

Astleitner, H. & Krumm, V. (1996). Dimensionen von Lehrverhalten: Faktorenstrukturen 1. und 2. Ordnung mit Kreuzvalidierung. *Empirische Pädagogik, 10* (1), 7–26.

Bachman, L. F. (2003). Commentary: Constructing an Assessment Use Argument and Supporting Claims about Test Taker-Assessment Task Interactions in Evidence-Centered Assessment Design. *Measurement: Interdisciplinary Research and Perspectives, 1* (1), 63–65.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly*, *2* (1), 1–34.

Bachman, L. F. (2015). Justifying the Use of Language Assessments: Linking Test Performance with Consequences. *JLTA Journal*, *18*, 3–22.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice. Developing language assessment and justifying their use in the real world* (Oxford applied linguistics). Oxford: Oxford University Press.

Backhaus, K. (2008). *Multivariate Analysemethoden. Eine Anwendungsorientierte Einführung* (12. Aufl.). Berlin: Springer.

Basler, H. D., Bolm, G., Dickescheid, T. & Herda, C. (1995). Marburger Fragebogen zur Akzeptanz der Lehre. *Diagnostica*, *41* (1), 62–79.

Baumert, B. & May, D. (2013). Constructive Alignment als didaktisches Konzept, *journal hochschuldidaktik*, *24* (1-2), 23–27.

Benson, J. (1998). Developing a Strong Program of Construct Validation: A Test Anxiety Example. *Educational Measurement: Issues and Practice, 17* (1), 10–27.

Bertrams, A. & Dickhäuser, O. (2009). Messung dispositioneller Selbstkontroll-Kapazität. Eine deutsche Adaptation der Kurzform der Self-Control Scale (SCS-K-D). *Diagnostica*, *55* (1), 2–10.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32* (3), 347–364.

Biggs, J. B. (1993). From Theory to Practice: A Cognitive Systems Approach. *Higher Education Research and Development*, *12* (1), 73–85.

Bollen, K. A. & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16* (3), 265–284.

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

Braun, E. (2007). *Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen (BEvaKomp)*. Göttingen: Vandenhoeck & Ruprecht unipress.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (Methodology in the social sciences). New York: Guilford Press.

Browne, W. J. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1* (3), 473–514.

Bühner, M. (2012). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erweiterte Aufl.). München: Pearson Studium.

Bundesregierung der Bundesrepublik Deutschland. (1998). *Viertes Gesetz zur Änderung des Hochschulrahmengesetzes*. HRG. Bundesgesetzblatt Teil I.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56* (2), 81–105.

Centra, J. A. (2003). Will teachers receive higher teacher evaluations by giving higher grades and less course work? *Research in Higher Education, 44* (5), 496–517.

Chung, Y., Rabe-Hesketh, Dorie, V., Gelman, A. & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78* (4), 685–709.

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies. *Review of Educational Research, 51* (3), 281–309.

Cohen, P. A. (1987, April). *A Critical Analysis and Reanalysis of the Multisection Validity Meta-Analysis*. Annual Meeting of the American Educational Research Association, Washington, DC.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Hrsg.), *Test validity* (S. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52* (4), 281–302.

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Hrsg.), *Educational Measurement* (2., überarb. Aufl., S. 443–507). Washington, DC.

Cronbach, L. J. (1972). *The Dependability of Behavioral Measurements. Theory of Generalizability for Scores and Profiles*. New York: Wiley.

D'Apollonia, S. & Abrami, P. C. (1996). *Variables moderating the validity of student ratings of instruction: A meta-analyses*. 77th. annual meeting of the American Educational Research Association, New York.

D'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52* (11), 1198–1208.

Die Europäischen Bildungsminister. (1999, 19. September). *Der Europäische Hochschulraum. Gemeinsame Erklärung der Europäischen Bildungsminister*. Zugriff am 29.09.2017. Verfügbar unter https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-03-Studium/02-03-01-Studium-Studienreform/Bologna_Dokumente/Bologna_1999.pdf

Diehl, J. M. & Kohr, H. U. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. *Psychologie in Erziehung und Unterricht, 24* (2), 61–75.

Dunnette, M. D. & Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology, 30* (1), 477–525.

Eberl, M. (2004). *Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des reflektiven Modells* (Schriften zur empirischen Forschung und quantitativen Unternehmensplanung). München: Ludwig-Maximilians-Universität München.

Eggert, A. & Fassott, G. (2003). *Zur Verwendung formativer und reflektiver Indikatoren in Strukturgleichungsmodellen: Ergebnisse einer Metaanalyse und Anwendungsempfehlungen* (Kaiserslauterer Schriftenreihe Marketing Nr. 23). Kaiserslautern: Universität Kaiserslautern, Lehrstuhl für Marketing.

Eisend, M. (2007). *Methodische Grundlagen und Anwendungen der Generalisierbarkeitstheorie in der betriebswirtschaftlichen Forschung* (Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin, Betriebswirtschaftliche Reihe, 2007/4). Berlin: Freie Universität Berlin, Fachbereich Wirtschaftswissenschaften.

Electric Paper - Gesellschaft für Softwarelösungen (Hrsg.). (2004). *EvaSys. Beispielhafte Implementation eines Fragebogeninstruments (HILVE-2) mit Unterstützung von Normen und Beratungstexten*. Lüneburg.

Embretson, S. E. & Reise, S. (2000). *Item Response Theory for Psychologists* (Multivariate Applications Book Series, Bd. 4). Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Esser, H. (1994*). Lehrbericht der Fakultät für Sozialwissenschaften der Universität Mannheim. Ergebnisse des Studenten- und Lehrerhebung im Wintersemester 1993/94*. Mannheim.

Fabrigar, L. R. & Wegener, D. T. (2012). *Exploratory Factor Analysis* (Series in understanding statistics). Oxford: Oxford University Press.

Fachbereich Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität Frankfurt am Main. (2014, 30. September). *Ordnung des Fachbereichs Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität Frankfurt am Main für den Bachelorstudiengang Psychologie vom 2. Juli 2014* (Der Präsident der Johann Wolfgang Goethe-Universität Frankfurt am Main, Hrsg.) (UniReport). Frankfurt am Main. Zugriff am 29.09.2017. Verfügbar unter http://www.psychologie.uni-frankfurt.de/65568207/BA-Psychologie_V2014.pdf

Fachbereich Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität, Frankfurt am Main. (2013, 12. September). *Ordnung des Fachbereichs Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität, Frankfurt am Main für den Masterstudiengang Psychologie vom 11. Mai 2011 in der Fassung vom 23. Januar 2013* (Der Präsident der Johann Wolfgang Goethe-Universität Frankfurt am Main, Hrsg.) (UniReport). Frankfurt am Main. Zugriff am

29.09.2017. Verfügbar unter http://www.psychologie.uni-frankfurt.de/50081109/Masterordnung-2013.pdf

Feistauer, D. & Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education, 42* (8), 1263–1279.

Feldman, K. A. (1976). The superior college teacher. *Research in Higher Education, 5* (3), 243–288.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30* (2), 137–194.

Finn, A. & Kayande, U. (1997). Reliability Assessment and Optimization of Marketing Measurement. *Journal of Marketing Research, 34* (2), 262–275.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik. Mit Hinweisen zur Intervention* (2., überarbeitete und erweiterte). Göttingen: Hogrefe.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education, 9* (1), 69–91.

Gelman, A. (2005). *Why I don't use the term "fixed and random effects".* Zugriff am 29.09.2017. Verfügbar unter http://andrewgelman.com/2005/01/25/why_i_dont_use/

Gerbing, D. & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25* (2), 186–192.

Gillmore, G. M. (1978). The Generalizability of Student Ratings of Instruction: Estimation of the Teacher and Course Components. *Journal of Educational Measurement, 15* (1), 1–13.

Goethe Graduate Academy. (2013). Feedback. Evaluationsbogen. Frankfurt am Main.

Gollwitzer, M. & Scholtz, W. (2003). Das" Trierer Inventar zur Lehrveranstaltungsevaluation"(TRIL): Entwicklung und erste testtheoretische Erprobungen. In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation* (Neue Medien, Konzepte, Untersuchungsbefunde und Erfahrungen zur psychologischen Aus-, Fort- und Weiterbildung, IV, S. 114–128). Berlin: Deutscher Psychologen Verlag.

Green, D. M. (1994). What is Quality in Higher Education? Concepts, Policy and Practice. In D. M. Green (Hrsg.), *What is quality in higher education?* (S. 3–20). Buckingham: Society for Research into Higher Education & Open University Press.

Green, S. B., Lissitz, R. W. & Mulaik, S. A. (1977). Limitations of Coefficient alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement, 37* (4), 827–838.

Guion, R. M. (1980). On Trinitarian Doctrines of Validity. *Professional Psychology*, *11* (3), 385–398.

Hadfield, J. (2010). MCMC Methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, *33* (2), 1–22.

Hartig, J., Frey, A. & Jude, N. (2007). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 135–164). Heidelberg: Springer Medizin Verlag.

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *New developments and techniques in structural equation modeling* (S. 89–127). New York: Psychology Press.

Helmke, A. (2006). Was wissen wir über guten Unterricht? Wissenschaftliche Erkenntnisse zur Unterrichtsforschung und Konsequenzen für die Unterrichtsentwicklung. *Pädagogik*, *58* (2), 42–45.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (1., neubearbeitete Aufl.). Seelze-Velber: Klett-Kallmeyer.

Iblher, P., Zupanic, M., Hartel, C., Heinze, H., Schmucker, P. & Fischer, M. R. (2011). The Questionnaire "SFDP26-German": A reliable tool for evaluation of clinical teaching? *GMS Zeitschrift für medizinische Ausbildung*, *28* (2), Doc30.

Johann Wolfgang Goethe-Universität. (2011). EvaSys. Studierenden-Fragebogen zur Lehrveranstaltungsevaluation. Frankfurt am Main.

Jonkisz, E. & Moosbrugger, H. (2007). Planung und Entwicklung von psychologischen Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–72). Heidelberg: Springer Medizin Verlag.

Junior, Fernando de Jesus Moreira, Zanella, A., Lopes, L. F. D. & Seidel, E. J. (2015). Student satisfaction evaluation through the Gradual Response Model of Item Response Theory [Avaliação da satisfação de alunos por meio do Modelo de Resposta Gradual da Teoria da Resposta ao Item]. *Ensaio: Avaliação e Políticas Públicas em Educação*, *23* (86), 129–158.

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, *112* (3), 527–535.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2* (3), 135–170.

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

Kaplan, D. & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling: A Multidisciplinary Journal*, *4* (1), 1–24.

Keil, W. (1975). *Kommunikation und Rezeption. Untersuchungen zur wissenschaftlichen Diskussion im Hochschulunterricht (*Arbeiten zur sozialwissenschaftlichen Psychologie, Bd. 3). Münster (Westfalen): Aschendorff.

Kember, D., Jenkins, W. & Chi Ng, K. (2004). Adult students' perceptions of good teaching as a function of their conceptions of learning—Part 2. Implications for the evaluation of teaching. *Studies in Continuing Education, 26* (1), 81–97.

Kember, D. & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education, 40* (1), 69–97.

Kieffer, J. M., Verrips, E. & Hoogstraten, J. (2009). Model specification in oral health-related quality of life research. *European Journal of Oral Sciences, 117* (5), 481–484.

Klauer, K. J. (1984). Kontentvalidität. *Diagnostica, 30* (1), 1–23.

Kleine, D. & Merkens, H. (1979). Überprüfung eines Fragebogens zur Beurteilung von Lehrveranstaltungen. *Psychologie in Erziehung und Unterricht, 26* (3), 149–153.

Kline, P. (1986). *A Handbook of Test Construction. Introduction to Psychometric Design*. London: Methuen & Co. Ltd.

Koch, E. (2004). *Gute Hochschullehre. Theoriebezogene Herleitung und empirische Erfassung relevanter Lehraspekte*. Hamburg: Verlag Dr. Kovac.

Köller, O. (2009). Evaluation pädagogisch-psychologischer Maßnahmen. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 333–352). Heidelberg: Springer Medizin Verlag.

Kramis, J. (1990). Bedeutsamkeit, Effizienz, Lernklima. Grundlegende Gütekriterien für Unterricht und didaktische Prinzipien. *Beiträge zur Lehrerbildung, 8* (3), 279–296.

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education, 4* (1), 1280256.

Kromrey, H. (1993). Studentische Vorlesungskritik. Empirische Daten und Konsequenzen für die Lehre. *Soziologie, 21* (1), 39–56.

Land Hessen. (2009). Hessisches Hochschulgesetz. HHG, HE. *Gesetz- und Verordnungsblatt für das Land Hessen* (22), 665–732. Zugriff am 29.09.2017. Verfügbar unter http://www.rv.hessenrecht.hessen.de/lexsoft/default/hessenrecht_rv.html?p1=0&eventSubmit_doNavigate=searchInSubtreeTOC&showdoccase=1&doc.hl=0&doc.id=jlr-HSchulGHE2010rahmen&doc.part=R&toc.poskey=

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

Lind, D. & Knoche, N. (2004). Testtheoretische Modelle und Verfahren bei PISA-2000-Mathematik. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von*

Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000 (S. 51–70). Wiesbaden: Verlag für Sozialwissenschaft.

Litzelman, D. K., Stratos, G. A., Marriott, D. J. & Skeff, K. M. (1998). Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. *Academic Medicine: Journal of the Association of American Medical Colleges, 73* (6), 688–695.

Lodico, M. G., Spaulding, D. T. & Voegtle, K. H. (2006). *Methods in Educational Research. From Theory to Practice.* San Francisco, CA: Jossey-Bass.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3* (3), 635–694.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Heidelberg: Springer.

Mackie, K. (1981). *The application of learning theory to adult teaching* (Adults, psychological and educational perspectives, Bd. 2). Nottingham: Department of Adult Education, University of Nottingham.

Mann, W. & Marshall, C. R. (2010). Building an Assessment Use Argument for Sign Language: the BSL Nonsense Sign Repetition Test. *International Journal of Bilingual Education and Bilingualism, 13* (2), 243–258.

Markus, K. A. & Borsboom, D. (2013). *Frontiers of Test Validity Theory. Measurement, Causation and Meaning* (Multivariate applications series). New York, N.Y.: Routledge / Taylor & Francis Group.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75* (1), 150–166.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11* (3), 253–387.

Marsh, H. W. (2007). Students' evaluations of university teaching dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Hrsg.), *The Scholarship of Teaching and Learning in Higher Education. An evidence-based perspective* (S. 319–383). Dordrecht: Springer.

Marsh, H. W. & Dunkin, M. J. (1996). Students' Evaluations of University Teaching: A Multidimensional Perspective. In R. P. Perry & J. C. Smart (Hrsg.), *Effective teaching in higher education: Research and Practice. Research and Practice* (S. 241–320). New York: Agathon Press.

Marsh, H. W. & Roche, L. A. (1992). The use of student evaluations of university teaching in different settings: The applicability paradigm. *Australian Journal of Education, 36* (3), 278–300.

Martin, F. & Booth, S. (1996). The learner's experience of learning. In D. R. Olson (Hrsg.), *The handbook of education and human development. New models of learning, teaching and schooling* (S. 534–564). Malden, Massachusetts: Blackwell.

McCrae, R. R. (2009). The five-factor model of personality traits: Consensus and controversy. In P. J. Corr & G. Matthews (Hrsg.), *The Cambridge Handbook of Personality Psychology* (S. 148–161). Cambridge: Cambridge University Press.

Messick, S. (1989a). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 19* (2), 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Hrsg.), *Educational Measurement* (3. Aufl., S. 13–103). New York, NY: American Council on Education & Macmillan.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A Brief Introduction to Evidence-centered Design*. Princeton: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1* (1), 3–62.

Moosbrugger, H. (2007). Klassische Testtheorie. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 99–112). Heidelberg: Springer Medizin Verlag.

Moosbrugger, H. & Kelava, A. (2007). Qualitätsanforderungen an einen psychologischen Test. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Heidelberg: Springer Medizin Verlag.

Moosbrugger, H. & Schweizer, K. (2002). Evaluationsforschung in der Psychologie. *Zeitschrift für Evaluation, 1* (1), 19–37.

Mulaik, S. A. & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal, 4* (3), 193–211.

Müller-Wolf, H.-M. (1977). *Lehrverhalten an der Hochschule: Dimensionen, Zusammenhänge, Trainingsmöglichkeiten*. München: Verlag Dokumentation.

Muthén & Bengt O. (1991). Multilevel Factor Analysis of Class and Student Achievement Components. *Journal of Educational Measurement, 28* (4), 338–354.

Myers, D. G. (2014). *Psychologie* (3. Aufl.). Heidelberg: Springer.

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods, 5* (2), 241–301.

Nunnally, J. C. (1967). *Psychometric Theory*. New York [u.a.]: McGraw-Hill, Inc.

Nunnally, J. C. (1978). *Psychometric Theory* (2. Überarb. Aufl.). New York [u.a.]: McGraw-Hill, Inc.

Onwuegbuzie, A. J., Daniel, L. G. & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student-teaching evaluations. *Quality and Quantity: International Journal of Methodology, 43* (2), 197–209.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D. & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44* (1), 113–160.

Ory, J. C. & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *27* (5), 27–44.

Paulitsch, M. A., Gerlach, F. M., Klingebiel, T. & Sennekamp, M. (2016). Auf dem Weg zum Dr. med. – Welche Unterstützung brauchen Promovierende der Medizin? Teil 2: Etablierung des Konzepts. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, *110-111*, 77–84.

Petter, S., Straub, D. & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, *31* (4), 623–656.

Piedmont, R. L. (2014). Factorial Validity. In A. C. Michalos (Hrsg.), *Encyclopedia of Quality of Life and Well-Being Research* (S. 2148–2149). Dordrecht: Springer Netherlands.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, *35* (1), 45–76.

Rammstedt, B. (2004). *Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung* (GESIS-How-to Nr. 12). Mannheim: Zentrum für Umfragen, Methoden und Analysen (ZUMA).

Rantanen, P. (2013). The Number of Feedbacks Needed for Reliable Evaluation: A Multilevel Analysis of the Reliability, Stability and Generalisability of Students' Evaluation of Teaching. *Assessment & Evaluation in Higher Education*, *38* (2), 224–239.

Rauch, W. A. & Moosbrugger, H. (2011). Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke (Hrsg.), *Methoden der psychologischen Diagnostik* (Enzyklopädie der Psychologie: Themenbereich B, Methodologie und Methoden. Serie 2, Psychologische Diagnostik, Bd. 2, Vollständige Neuausgabe, S. 1–87). Göttingen [u.a.]: Hogrefe, Verlag für Psychologie.

Rindermann, H. (2009). *Lehrevaluation* (Psychologie, Bd. 42, 2., leicht korrigierte Auflage). Landau: Empirische Pädagogik.

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation. A systematic approach*. Newbury Park, Calif.: Sage.

Schermelleh-Engel, K. & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 113–164). Heidelberg: Springer Medizin Verlag.

Schmitz, C. (2012). *LimeSurvey: An open source survey tool*. (LimeSurvey Project, Hrsg.) Hamburg, Germany. Zugriff am 29.09.2017. Verfügbar unter http://www.limesurvey.org.

Schwarz, N. (1996). *Cognition and communication. Judgmental biases, research methods, and the logic of conversation* (John M. MacEachran memorial lecture series). Mahwah, NJ: Erlbaum.

Searle, S. R., Casella, G. & McCulloch, C. (1992). *Variance Components* (Wiley Series in Probability and mathematical Statistics). New York, Chichester, Brisbane, Toronto, Singapore: John Wiley & Sons, Inc.

Seldin, P. (1993). How colleges evaluate professors, 1983 vs. 1993. *AAHE Bulletin, 12* (110), 6–8.

Sengewald, E. & Vetterlein, A. (2015). Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation. *Diagnostica, 61* (3), 116–123.

Sennekamp, M., Paulitsch, M. A., Broermann, M., Klingebiel, T. & Gerlach, F. M. (2016). Auf dem Weg zum Dr. med. – Welche Unterstützung brauchen Promovierende der Medizin? Teil 1: Bestandsaufnahme und Konzeptentwicklung. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 110-111*, 69–76.

Shavelson, R. J. & Webb, N. M. (1991). *A Primer on Generalizability Theory* (Measurement methods for the social sciences, Bd. 1). Newbury Park, Calif.: Sage Publications.

Sippel, S. (2014, September). *Dozenten-Evaluation im Medizinstudium: Psychometrische Charakteristika verfügbarer Messinstrumente*, Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA), Hamburg.

Skeff, K. M. (1988). Enhancing teaching effectiveness and vitality in the ambulatory setting. *Journal of General Internal Medicine, 3* (Supplement 2), 33.

Skeff, K. M., Stratos, G. A., Berman, J. & Bergen, M. R. (1992). Improving clinical teaching. Evaluation of a national dissemination program. *Archives of Internal Medicine, 152* (6), 1156–1161.

Müller-Wolf, H.M. & Fittkau, F. (1971). Lehrverhalten von Hochschullehrern und seine Bedeutung für Einstellungen und Verhalten von Studenten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 3*, 165–180.

Souvignier, E. & Gold, A. (2002). Fragebögen zur Lehrevaluation: Was können sie leisten? *Zeitschrift für Evaluation, 1* (2), 265–280.

Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research, 83* (4), 598–642.

Staufenbiel, T. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica, 46* (4), 169–181.

Stemmler, G., Bartussek, D., Hagemann, D. & Amelang, M. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (Kohlhammer Standards Psychologie, 7., vollständig überarbeitete Aufl.). Stuttgart: Kohlhammer.

Terzer, E., Hartig, J. & Upmeier zu Belzen, A. (2013). Systematische Konstruktion eines Tests zu Modellkompetenz im Biologieunterricht unter Berücksichtigung von Gütekriterien. *Zeitschrift für Didaktik der Naturwissenschaften*, *19* (1), 51–76.

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press.

Toland, M. D. (2005). A Multilevel Factor Analysis of Students' Evaluations of Teaching. *Educational and Psychological Measurement*, *65* (2), 272–296.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Wenhold, F., Elbe, A.-M. & Beckmann, J. (2009). Testgütekriterien des Fragebogens VKS zur Erfassung volitionaler Komponenten im Sport. *Zeitschrift für Sportpsychologie*, *16* (3), 91–103.

Westermann, R., Spies, K., Heise, E. & Wollburg-Claar, S. (1998). Bewertung von Lehrveranstaltungen und Studienbedingungen durch Studierende: Theorieorientierte Entwicklung von Fragebögen. *Empirische Pädagogik*, *12* (2), 133–166.

Winteler, A. (1974). *Determinanten der Wirksamkeit akademischer Lehrveranstaltungen* (Psychologia universalis, Bd. 30). Meisenheim am Glan: Hain.

Winteler, A. & Schmolck, P. (1979). Entwicklung und Validierung eines Schätzverfahrens zur Beurteilung von Lehrveranstaltungen. *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen*, *38* (2), 139–156.

Winteler, A. & Schmolck, P. (1983). Überprüfung eines Schätzverfahrens zur Beurteilung von Lehrveranstaltungen. *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen*, *42* (1), 56–79.

Wissenschaftsrat. (2008). *Empfehlungen zur Qualitätsverbesserung von Lehre und Studium.* Berlin

Zentrum für Qualitätssicherung und -entwicklung. (2011). *Aspekte guter Lehre an der Johannes Gutenberg-Universität Mainz* (Prof. Dr. Georg Krausch, Hrsg.). Verfügbar unter http://www.uni-mainz.de/lehre/Dateien/JGU_aspekte_guter_lehre.pdf

Anhang

Anhang A: Das Lehrevaluationsinventar des Frankfurter Promotionskollegs am Fachbereich Medizin

Anhang B: Das Lehrevaluationsinventar der Goethe Graduate Academy (GRADE)

Anhang C: Der für alle Veranstaltungen der Goethe-Universität gültige Teil des Lehrevaluationsinventars

Anhang D: Die Syntax der Varianzkomponentenschätzung anhand des r-Pakets MCMCglmm für das Item „Die Lernziele der Verantaltlungs waren klar erkennbar“

Anhang E: Das Ergebnis der Varianzkomponentenschätzung im Detail mit dem Mittelwert der simulierten Varianzkomponenten und entsprechendem Glaubwürdigkeitsintervall (GI)

Anhang F: Versicherung und Erklärung

Anhang G: Detaillierte Darstellung der Eigenleistung

Anhang A: Das Lehrevaluationsinventar des Frankfurter Promotionskollegs am Fachbereich Medizin

UNIVERSITÄT FRANKFURT AM MAIN

Modul-Evaluation des Promotionskollegs

Liebe Doktorandinnen und Doktoranden,
wir sind ständig bemüht, die Qualität unserer Unterrichtsveranstaltungen zu optimieren. Dazu benötigen wir Ihre Mithilfe. Bitte geben Sie uns Feedback zur heutigen Veranstaltung. Ihre Daten werden selbstverständlich anonym ausgewertet, dazu hilft uns der folgende Code:

1. **Dritter** Buchstabe des **Vornamens** Ihres **Vaters**
2. **Vierter** Buchstabe Ihres **Geburtsortes**
3. **Zweiter** Buchstabe des **Geburtsnamens** Ihrer **Mutter**
4. **Vierter** Buchstabe (bzw. letzter) des **eigenen Vornamens**
5. **Fünfte** Ziffer Ihrer **Matrikel-Nummer**

Code

1	2	3	4	5

Geschlecht: ☐ weiblich ☐ männlich **Geburtsjahr:** ☐☐☐☐

Thema der Veranstaltung: _____

Dozent(in): _____ **Datum:** _____

Stand Ihrer Dissertation:

Bitte kreuzen Sie an, in welcher Phase Ihrer Dissertation Sie sich momentan befinden (Mehrfachantworten sind möglich):

☐ noch nicht begonnen ☐ Literaturrecherche ☐ Daten-Sammlung ☐ Daten-Auswertung ☐ Schreiben

Bitte kreuzen Sie an, welche Phase(n) Sie schon abgeschlossen haben:

☐ keine ☐ Literaturrecherche ☐ Daten-Sammlung ☐ Daten-Auswertung ☐ Schreiben

Bewertung des Kurses: Bitte bewerten Sie den Kurs, indem Sie jeweils ein Feld pro Aussage ankreuzen.

	stimme zu	**stimme eher zu**	**stimme eher nicht zu**	**stimme nicht zu**
Die Lernziele der Veranstaltung waren klar erkennbar	☐	☐	☐	☐
Ich hatte die Möglichkeit, mich aktiv an der Veranstaltung zu beteiligen	☐	☐	☐	☐
Die Arbeitsatmosphäre war konstruktiv	☐	☐	☐	☐
Die Erklärungen des Dozenten waren verständlich	☐	☐	☐	☐
Die Durchführung durch den Dozenten war motivierend	☐	☐	☐	☐
Die verwendeten Unterrichtsmaterialien waren angemessen	☐	☐	☐	☐
Ich hatte ausreichende Vorkenntnisse, um der Veranstaltung zu folgen	☐	☐	☐	☐

UNIVERSITÄT FRANKFURT AM MAIN

Modul-Evaluation des Promotionskollegs

	stimme zu	stimme eher zu	stimme eher nicht zu	stimme nicht zu
Ich glaube, die heute erlernten Inhalte in meiner Dissertation umsetzen zu können	☐	☐	☐	☐
Die Inhalte der heutigen Veranstaltung sind für mich persönlich relevant	☐	☐	☐	☐
Die Veranstaltung war gut organisiert	☐	☐	☐	☐
Ich würde diese Veranstaltung anderen Doktoranden weiterempfehlen	☐	☐	☐	☐

Der Zeitrahmen der Veranstaltung war	☐ zu kurz	☐ genau richtig	☐ zu lang
Die Stoffmenge der Veranstaltung war	☐ zu wenig	☐ genau richtig	☐ zu viel
Das Unterrichtstempo des Dozenten war	☐ zu langsam	☐ genau richtig	☐ zu schnell

Ich habe Interesse an einem weiterführenden Aufbaukurs zu diesem Thema ☐ ja ☐ nein

Gesamtnote für diese Veranstaltung: ☐ *(in Schulnoten: 1= sehr gut bis 6= ungenügend)*

Folgende Themen würde ich zusätzlich in die Veranstaltung mit aufnehmen oder ausführlicher behandeln:

Folgende Themen würde ich aus der Veranstaltung kürzen oder streichen:

Ich habe folgende Verbesserungsvorschläge für die Unterrichtsmaterialien und -methoden:

Das fand ich besonders gut:

Das würde ich ändern:

Vielen Dank für Ihre Mithilfe und viel Erfolg für Ihre Dissertation!
Ihr Team des Promotionskollegs

Anhang B: Der für alle Veranstaltungen der Goethe-Universität gültige Teil des Lehrevaluationsinventars

EvaSys	Studierenden-Fragebogen Promotionskolleg 2011	

Johann Wolfgang Goethe-Universität
FB 16 Medizin 2011

Markieren Sie so: ☐☒☐☐☐ Bitte verwenden Sie einen Kugelschreiber oder nicht zu starken Filzstift. Dieser Fragebogen wird maschinell erfasst.
Korrektur: ☐■☐☒☐ Bitte beachten Sie im Interesse einer optimalen Datenerfassung die Hinweise zum Ausfüllen.

Liebe Veranstaltungsteilnehmerin, lieber Veranstaltungsteilnehmer,

Im folgenden finden Sie eine Reihe von Aussagen über die Lehrveranstaltung, in der Sie diesen Fragebogen erhalten haben. Schätzen Sie bitte durch Ankreuzen einer Antwortalternative ein, in welchem Ausmaß diese Aussagen Ihrer Meinung nach auf die jeweilige Lehrveranstaltung zutreffen. Sie haben hierzu sechs Antwortmöglichkeiten von 1 ("trifft nicht zu") bis 6 ("trifft zu"). Bearbeiten Sie die einzelnen Aussagen bitte zügig, aber nicht ohne sie gründlich gelesen zu haben.
Vielen Dank für Ihre Mitarbeit.. Dr. A. Syed Ali ; Dekanat/M. Sennekamp; Institut für Allgemeinmedizin

1. Allgemeine Aussagen zur Lehrveranstaltung

1.1	Der Besuch der Veranstaltung führt zu einem spürbaren Wissenszuwachs.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.2	Der in der Veranstaltung vermittelte Stoff ist gut strukturiert.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.3	In der Veranstaltung werden ausreichend Hilfsmittel zur Aneignung des Lehrstoffs (Skripte, Lehrtexte, Literaturlisten etc.) angeboten.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.4	Das Tempo der Veranstaltung ist angemessen	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.5	In der Veranstaltung werden auch schwierige Inhalte verständlich erklärt.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.6	Der Veranstalter / die Veranstalterin geht auf Fragen der Teilnehmer/-innen angemessen ein.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.7	In der Veranstaltung werden Medien (Tafel, Folien, PowerPoint-Screens etc.) in geeigneter Weise eingesetzt.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.8	In der Veranstaltung wird ein guter Überblick über das behandelte Stoffgebiet vermittelt.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.9	In der Veranstaltung sind inhaltliche Zusammenhänge ("roter Faden") deutlich erkennbar.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.10	Aktuelle Fragestellungen werden in die Veranstaltung angemessen integriert.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.11	Eine selbständige und aktive Auseinandersetzung mit den Lerninhalten wird in der Veranstaltung gefördert.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu
1.12	In der Lehrveranstaltung herrscht ein konstruktives, positives Klima.	trifft nicht zu	☐	☐	☐	☐	☐	☐	trifft zu

Anhang C: Das Lehrevaluationsinventar der Goethe Graduate Academy (GRADE)

Feedback

GRADE
GOETHE GRADUATE ACADEMY

Workshop: xxx
Date: xxx, 2017
Lecturer: xxx

	very much / vollkommen	rather yes / eher ja	partly / teils teils	rather no / eher nicht	not at all / gar nicht
The workshop met my expectations. / Der Workshop hat meine Erwartungen erfüllt.	1 ❑	2 ❑	3 ❑	4 ❑	5 ❑
The workshop was clearly structured. / Der Workshop war klar strukturiert.	1 ❑	2 ❑	3 ❑	4 ❑	5 ❑
The lecturer presented the contents clearly and coherently. / Der/die Trainer/in hat die Inhalte klar und anschaulich vermittelt.	1 ❑	2 ❑	3 ❑	4 ❑	5 ❑
I would recommend the workshop. / Ich würde den Workshop weiterempfehlen.	1 ❑	2 ❑	3 ❑	4 ❑	5 ❑
The content of the workshop was helpful. / Der Inhalt des Workshops war hilfreich.	1 ❑	2 ❑	3 ❑	4 ❑	5 ❑

For what? / Wofür?

- ❑ Dissertation / Promotion
- ❑ Career / Karriere
- ❑ Personal Development / Entwicklung der Persönlichkeit
- ❑ Other, please explain /Sonstiges, bitte führen Sie aus:

	definitely too long deutlich zu lang	slightly too long / etwas zu lang	appropriate / angemessen	slightly too short / etwas zu kurz	definitely too short / deutlich zu kurz
The duration of the workshop was appropriate. / Die Länge des Workshops war angemessen.	5 ❑	3 ❑	1 ❑	-3 ❑	-5 ❑

How did you find out about this workshop? / Wie haben Sie von diesem Workshop erfahren?

- ❑ Through my supervisor / Durch meine/n Betreuer/in
- ❑ Through fellow doctoral candidates / Durch andere Promovierende
- ❑ Through the GRADE workshop brochure / Durch die GRADE-Workshopbroschüre
- ❑ Through an e-mail from GRADE / Durch eine E-Mail von GRADE
- ❑ Through poster or flyer / Durch Poster oder Flyer
- ❑ Through GRADE Homepage / Durch die GRADE-Webseite
- ❑ Other, please explain / Andere, bitte führen Sie aus:_____

Which themes of this workshop did you specifically like?
Welche Workshopinhalte haben Sie besonders angesprochen?

Are there any themes or methodologies missing in the workshop?
Fehlten in diesem Workshop bestimmte Inhalte oder Methoden?

Further Comments:
Weitere Kommentare:

What other workshops/contents should be offered by GRADE?
Welche weiteren Themen/Inhalte sollte GRADE anbieten?

Thank you for your feedback!

Danke für Ihre Rückmeldung!

Anhang D: Die Syntax der Varianzkomponentenschätzung anhand des r-Pakets MCMCglmm für das Item „Die Lernziele der Veranstaltung waren klar erkennbar“

```
lernziele <- MCMCglmm(Lernziele ~ 1, random = ~ VPID + Thema + Dozent,
data=pk_daten, nitt=10000, thin=10, burnin=5000)
```

Anhang E: Das Ergebnis der Varianzkomponentenschätzung im Detail mit dem Mittelwert der simulierten Varianzkomponenten und entsprechendem Glaubwürdigkeitsintervall (GI)

	Lernziele		*Beteiligung*		*Atmosphäre*	
	Mittelwert	*GI (95%)*	*Mittelwert*	*GI (95%)*	*Mittelwert*	*GI (95%)*
Teilnehmer	0.027	0.018 - 0.037	0.042	0.03 - 0.055	0.041	0.032 - 0.051
Thema	0.019	0.001 - 0.05	0.000	0.000 - 0.001	0.000	0.000 - 0.001
Dozent	0.044	0.018 - 0.073	0.145	0.077 - 0.244	0.077	0.036 - 0.127
Residuum	0.214	0.2 - 0.229	0.254	0.239 - 0.271	0.199	0.187 - 0.211
	Verständlichkeit		*Motivation*		*Material*	
	Mittelwert	*GI (95%)*	*Mittelwert*	*GI (95%)*	*Mittelwert*	*GI (95%)*
Teilnehmer	0.04	0.029 - 0.052	0.055	0.04 - 0.072	0.05	0.036 - 0.064
Thema	0.000	0.000 - 0.000	0.000	0.000 - 0.001	0.004	0.000 - 0.013
Dozent	0.075	0.041 - 0.125	0.118	0.061 - 0.196	0.05	0.018 - 0.084
Residuum	0.199	0.188 - 0.21	0.302	0.281 - 0.321	0.248	0.233 - 0.266
	Vorkenntnisse		*Umsetzbarkeit*		*Relevanz*	
	Mittelwert	*GI (95%)*	*Mittelwert*	*GI (95%)*	*Mittelwert*	*GI (95%)*
Teilnehmer	0.183	0.149 - 0.223	0.071	0.055 - 0.096	0.078	0.058 - 0.096
Thema	0.047	0.01 - 0.114	0.084	0.013 - 0.217	0.088	0.015 - 0.227
Dozent	0.045	0.014 - 0.084	0.049	0.023 0.087	0.043	0.017 - 0.079
Residuum	0.413	0.383 - 0.436	0.311	0.289 - 0.331	0.304	0.285 - 0.323
	Organisation		*Empfehlung*		*Gesamtnote*	
	Mittelwert	*GI (95%)*	*Mittelwert*	*GI (95%)*	*Mittelwert*	*GI (95%)*
Teilnehmer	0.034	0.023 - 0.045	0.057	0.039 - 0.073	0.094	0.072 - 0.123
Thema	0.001	0.000 - 0.005	0.006	0.000 - 0.034	0.000	0.000 - 0.001
Dozent	0.054	0.026 - 0.092	0.113	0.049 - 0.187	0.2	0.097 - 0.327
Residuum	0.243	0.227 - 0.257	0.328	0.309 - 0.349	0.411	0.385 - 0.436

Anhang F: Versicherung und Erklärung

Versicherung

Ich erkläre hiermit, dass ich die vorgelegte Dissertation über „Die Validität der Interpretationen studentischer Lehrevaluationsergebnisse: Eine exemplarische Anwendung des argumentationsbasierten Ansatzes“ selbstständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe, insbesondere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind.

Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis beachtet und nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben.

Frankfurt am Main, den 09. Oktober 2017 ….....

(Unterschrift)

Erklärung

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung im philosophischen Bereich unterzogen habe und dass mir die Promotionsordnung bekannt ist.

Frankfurt am Main, den 09. Oktober 2017 ….....

(Unterschrift)

Anhang G: Detaillierte Darstellung der Eigenleistung

Erklärung zur Eigenleistung des Doktoranden bei der Dissertation

Meine Eigenleistung an der Entwicklung der Fragestellung, Design, Durchführung und Auswertung der empirischen Studie beträgt 100%.

Frankfurt am Main, den 09. Oktober 2017 ….....

(Unterschrift Doktorand)

Frankfurt am Main, den 09. Oktober 2017 ….....

(Unterschrift Betreuer)

Danksagung

Ich möchte folgenden Personen für ihre Hilfsbereitschaft bei der Beantwortung verschiedener Fragen, die in dieser Dissertation aufgekommen sind, danken: Christine Aichele für die Diskussion darüber, was eine Validitäts-Argumentation ausmacht, Prof. Dr. Gilberto Alves für die Übersetzungen aus dem Portugiesischen, Dr. phil. Edith Braun für die Bereitstellung von Literatur zu Lehrevaluation, Dr. phil. Anna Praetorius für das Verständnis der Generalisierbarkeitstheorie und Dr. phil. Alexander Robitzsch für die Beratung zur Auswahl eines angemessenen Schätzalgorithmus. Ganz herzlichen Dank geht an Dr. phil. Sabine Fabriz und Dr. phil. Miriam Hansen für Ihre weiterführende Hilfe.

Besonderen Dank geht an Prof. Dr. Johannes Hartig für die zuverlässige und kompetente Betreuung in angenehmer Atmosphäre sowie an die Leiterin des Frankfurter Promotionskollegs am Fachbereich Medizin Dr. phil. Monika Sennekamp, die zu jedem Zeitpunkt bereit war, das Gelingen dieses Promotionsprojektes zu unterstützen.