Personalpsychologie

Band 3

Evaluationsfragebogen zur **E**rfassung **s**tudentischer **L**ehrurteile (EEsL)

Frank Heber

2016

Zusammenfassung

In dieser Arbeit wurden vorliegende Fragebögen zur Evaluation von Lehrveranstaltungen hinsichtlich ihrer Eignung zum Einsatz im Rahmen der regelmäßigen studentischen Lehrveranstaltungsevaluation an der Hochschule Hannover – Fakultät IV – Abteilung Betriebswirtschaft geprüft. Ziel war es, einen ökonomischen und methodisch überzeugenden Fragebogen zu identifizieren. Die Sichtung der einschlägigen Literatur sowie der vorliegenden Fragebögen zeigte, dass keiner dieser Fragebögen zur Evaluation von Lehrveranstaltungen alle relevanten Kriterien vollständig erfüllte. Deshalb wurden alle Items der recherchierten Fragebögen für die Verwendung innerhalb eines neuen Fragebogens inhaltlich gruppiert und überprüft. Im Ergebnis steht ein ökonomisch und unter Berücksichtigung methodisch aktueller Befunde konzipierter Fragebogen zur Verfügung, der sowohl für den Einsatz in der regelmäßigen studentischen Evaluation von Lehrveranstaltungen einer Hochschule geeignet ist als auch für den Einsatz in Fort- und Weiterbildungsveranstaltungen.

Heber, Frank
Evaluationsfragebogen zur Erfassung studentischer Lehrurteile (EEsL). – Hannover : Hochschule Hannover, 2016
(Personalpsychologie; Band 3).
ISSN: 2199-9759

Weitere Schriften aus der Reihe *Personalpsychologie* finden Sie unter:
http://serwiss.bib.hs-hannover.de/solrsearch/index/search/searchtype/series/id/7
Die Schriftenreihe *Personalpsychologie* enthält Schriften zu Personaldiagnostik, Personalentwicklung und Personalführung

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; *detaillierte bibliografische Daten* sind im Internet über http://dnb.ddb.de abrufbar.

Impressum

Herausgegeben von Prof. Dr. Sven Litzcke,
Professur für Human Resource Management und Wirtschaftspsychologie
c/o Hochschule Hannover, Fakultät IV - Abteilung Betriebswirtschaft
Ricklinger Stadtweg 120
30459 Hannover

Inhalt

1 Einleitung

Mit dieser Arbeit soll die Einführung eines neuen Fragebogens zur Evaluation von Lehrveranstaltungen an der Hochschule Hannover, Fakultät IV – Wirtschaft und Informatik, Abteilung Betriebswirtschaft (im Folgenden nur noch Abteilung Betriebswirtschaft genannt) unterstützt werden. Mindestens einmal im Jahr müssen gemäß Paragraph 5 Absatz 1 Satz 1 des Niedersächsischen Hochschulgesetzes (kurz: NHG) Lehrveranstaltungen durch Studierende evaluiert werden. Entscheidend für die geplante Ablösung des derzeit verwendeten Fragebogens zur Evaluation von Lehrveranstaltungen waren die starken empirischen Zusammenhänge in gemessenen Dimensionen, die theoretisch-inhaltlich unabhängig sind. Eine Analyse der Ergebnisse der Lehrveranstaltungsevaluation des Sommersemesters 2016 ergab hohe Korrelationen der drei theoretisch-inhaltlich unabhängigen Bereiche *Inhalt*, *Dozent* und *Didaktik* von r=.82 bis r=.90. Im Wintersemester 2015/2016 (r=.78 bis r=.86) und im Sommersemester 2015 (r=.86 bis r=.89) wurden diese Ergebnisse bestätigt. Die Höhe der Korrelationen weist darauf hin, dass Studierende nur unzureichend zwischen den verschiedenen abgefragten Bereichen unterscheiden und lediglich ein Globalurteil über eine Lehrveranstaltung in ihren Antworten abbilden. Angesichts dessen, dass Studierende in ihren Antworten nicht erkennbar zwischen den drei Bereichen *Inhalt*, *Dozent* und *Didaktik* unterscheiden, ist der derzeit verwendete Fragebogen zu umfangreich und verbraucht unnötig Ressourcen.

Zudem weist der derzeit verwendete Fragebogen gravierende methodische Mängel auf, die exemplarisch verdeutlich werden sollen. Beispielsweise erfasst die Frage „Wie oft haben Sie sich durch Fragen und/oder Antworten beteiligt?“ inhaltlich zwei unterschiedliche Aspekte. Was soll ein Studierender antworten, der sehr oft Antworten gegeben aber wenig Fragen gestellt hat, und was sagt eine Antwort auf eine solche Frage aus? Ein vergleichbarer Mangel liegt bei dem Item „Die zur Verfügung gestellten Materialien sind hilfreich und aktuell.“ vor, das ebenfalls zwei Aspekte erfasst und zudem durch Studierende kaum verlässlich zu beurteilen ist, wie ein Gespräch mit Vertretern des Fachschaftsrats Wirtschaft belegt. Abgesehen von der zeitgleichen Abfrage zweier Aspekte, ist die Frage auch inhaltlich wenig überzeugend. Mitunter können gerade im Grundlagenbereich gelehrte Inhalte Jahrzehnte alt sein, aber noch immer einschlägig und aktuell, was aber nur verlässlich beurteilt werden kann, wenn bereits ein Überblick über ein Themengebiet besteht. Davon kann jedoch bei Studierenden niedriger Semester regelmäßig nicht ausgegangen werden. Sieht man von den methodischen Besonderheiten retrospektiver Fragen ab (Blossfeld, 2010: 998; Saris/Gallhofer, 2007: 85), ist die Frage nach dem durchschnittlichen wöchentlichen Arbeitsaufwand insofern heikel, als in der Abteilung Betriebswirtschaft die Evaluation von Lehrveranstaltungen zu zwei Erhebungszeitpunkten durchgeführt werden kann. Zur Auswahl stehen ein Zeitpunkt in der Mitte des Semesters und ein Zeitpunkt gegen Ende des Semesters. In der Mitte des Semesters ist der durchschnittliche wöchentliche Arbeitsaufwand für das gesamte Semester aber kaum verlässlich zu beurteilen, insbesondere dann nicht, wenn Studierende erst gegen Endes des Semesters für eine Prüfung zu lernen beginnen. Darüber hinaus wird aus der Frage nach dem durchschnittlichen wöchentlichen Arbeitsaufwand nicht ersichtlich, ob sich die Frage auf die aktuell zu bewertende Lehrveranstaltung bezieht oder allgemein auf alle Lehrveranstaltungen, die ein Studierender besucht. Ferner weisen einzelne Ausprägungen des geschlossenen Antwortformats bei der Frage nach dem durchschnittlichen wöchentlichen Arbeitsaufwand Überschneidungen untereinander auf. Es bleibt unklar, was ein Studierender antworten soll, der sich seiner Ansicht nach in der Schnittmenge zweier Antwortkategorien bewegt.

Abgesehen von methodischen Mängeln bei einzelnen Fragen, ist zudem das variierende Antwortformat des derzeit verwendeten Fragebogens zu kritisieren. Einige der Fragen beinhalten

ein vierstufiges vollständig beschriftetes Antwortformat, während andere Fragen ein sechsstufiges lediglich endpunktskaliertes Antwortformat aufweisen. Auf der Rückseite des Fragebogens fehlt zudem die Beschriftung des sechsstufigen Antwortformats. Darüber hinaus ließe sich innerhalb der Instruktionen auf eine mit einem Ausrufezeichen versehene Botschaft verzichten, die vermutlich instruktiv gemeint ist, die aber auch als interessengeleitet wahrgenommen werden kann. Es bleibt diesbezüglich unklar, weshalb die durchgängige Verwendung der weiblichen Form (die Dozentin), der männlichen Form (beispielsweise der Dozent) vorzuziehen sei und zu einer „Vereinfachung" führen soll. Wenn schon innerhalb eines Fragebogens zur Evaluation von Lehrveranstaltungen derart argumentiert wird, wäre ein sprachlich neutralerer Begriff wie *Lehrperson* vorzuziehen, der tatsächlich beiden biologischen Geschlechtern gerecht würde.

Aus methodischer Sicht ist der derzeit verwendete Fragebogen zur Evaluation von Lehrveranstaltungen der Abteilung Betriebswirtschaft nicht geeignet und sollte abgelöst werden, da eine Messung methodischen Mindeststandards sowie psychometrischen Gütekriterien nicht gerecht wird und Befunde folglich nur eingeschränkt inhaltlich interpretiert werden sollten. Aus den dargelegten Gründen empfiehlt sich die Auswahl, oder sofern sich kein geeigneter Fragebogen finden lässt, die Konzeption eines neuen Fragebogens zur Evaluation von Lehrveranstaltungen durch Studierende. Etwas überraschend war die Vielzahl der methodischen Mängel des derzeit verwendeten Fragebogens.

Wie im Verlauf der Arbeit gezeigt wird, werden an vielen Universitäten und Fachhochschulen (im Folgenden nur noch Hochschulen genannt) methodisch wenig überzeugende Fragebögen eingesetzt. Wenn selbst an Hochschulen Fragebögen mit methodisch unzureichendem Fundament verwendet werden, obwohl die fachwissenschaftliche Expertise zur Erstellung von Fragebögen – im Vergleich zur beruflichen Praxis – an einer Hochschule eigentlich vorhanden sein sollte, ist für die berufliche Praxis von methodisch ähnlich zweifelhaften Fragebögen auszugehen. Für die berufliche Praxis wird daher ein methodisch ähnliches Niveau der Fragebögen zur Beurteilung von Lehrenden im Rahmen von Fort- und Weiterbildungsveranstaltungen vermutet. Ein Nebenziel war daher einen Fragebogen zu finden, beziehungsweise zu entwickeln, der nicht nur für den Einsatz in Lehrveranstaltungen an Hochschulen geeignet ist, sondern möglichst auch in der Fort- und Weiterbildung eingesetzt werden kann. Interessant wäre der Einsatz eines methodisch einwandfrei konzipierten Fragebogens in der beruflichen Praxis beispielsweise bei Fort- und Weiterbildungsmaßnahmen mit Frontalunterricht und Lehrendem, da bekannt ist, dass Lehrende die Trainingseffektivität beeinflussen (Höft 2014: 1121). Ebenso kann die Evaluation zur Legitimation personalpsychologischer Arbeit allgemein beitragen (Höft, 2014: 1081), denn gemäß Statistischem Bundesamt (2013: 49) entstanden in 2010 je Teilnehmenden an Lehrveranstaltungen zur Fort- und Weiterbildung Gesamtkosten in Höhe von rund 1.500 Euro. Seyda und Werner (2014: 6) berichten für das Jahr 2013 Gesamtkosten in Höhe von rund 1.100 Euro, berücksichtigen in ihren Berechnungen aber keine Auszubildenden. Beinhaltet sind in den Gesamtkosten die jeweiligen Kosten der Lehrveranstaltung je Teilnehmenden und dessen Personalausfallkosten. Ein methodisch gut konzipierter Fragebogen zur Evaluation von Lehrveranstaltungen der Fort- und Weiterbildung ist nur ein Aspekt im Rahmen einer umfassenden Evaluation von Personalentwicklung. Würde eine Lehrveranstaltung der Fort- und Weiterbildung gut beurteilt, böte sich eine Rechtfertigungsgrundlage für die entstehenden Kosten. Bis dato nutzen lediglich etwas mehr als die Hälfte der Unternehmen (rund 52 Prozent), die Möglichkeit der Evaluation von Fort- und Weiterbildungsaktivitäten (Statistisches Bundesamt, 2013: 81), und das, obwohl das gesamtwirtschaftliche Investitionsvolumen in Deutschland auf mehr als 33 Milliarden Euro geschätzt wird (Seyda/Werner, 2014: 7).

Im Folgenden werden theoretische Grundlagen der Evaluation von Lehrveranstaltungen erläutert. Darüber hinaus werden empirische Befunde zur Durchführung einer Lehrveranstaltungsevaluation und zu möglichen Verzerrungsvariablen präsentiert. Abgerundet wird Kapitel 2 durch Fragebögen, die im Rahmen dieser Arbeit recherchiert und geprüft wurden.

2 Theorie

Im Folgenden wird der Begriff *Lehrveranstaltungsevaluation* definiert und es wird geprüft, welche Durchführungsart (Online versus Papier) empfehlenswert ist, zu welchem Zeitpunkt die Evaluation von Lehrveranstaltungen am sinnvollsten durchgeführt werden sollte und welche Verzerrungsvariablen auftreten können. Im letzten Abschnitt des Kapitels werden die Fragebögen aufgeführt, die hinsichtlich ihres Einsatzes als neuer Fragebogen recherchiert und geprüft wurden.

2.1 Evaluation von Lehrveranstaltungen

Die Evaluation von Lehrveranstaltungen durch Studierende ist nur ein Aspekt verschiedener Evaluationsprozesse, die im Rahmen einer umfassenden Qualitätssicherung an Hochschulen durchgeführt werden. Auf Makro-Ebene beispielsweise lässt sich der Übergang ins Berufsleben evaluieren, auf Meso-Ebene beispielsweise Studiengänge und auf Mikro-Ebene beispielsweise die Evaluation einzelner Lehrveranstaltungen (el Hage, 1996: 2; Stratmann, 2016: 148).

In Niedersachsen ist die Verbindlichkeit zur Evaluation von Lehrveranstaltungen gesetzlich vorgeschrieben. Gemäß Paragraph 5 Absatz 1 Satz 1 des NHG muss Studierenden mindestens einmal im Jahr die Möglichkeit zur Evaluation von Lehrveranstaltungen eingeräumt werden. In Niedersachsen waren zum Sommersemester 2016 rund 190.000 Studierende an staatlichen Hochschulen eingeschrieben (Landesamt für Statistik Niedersachsen, 2016). Da zentrale Ergebnisse der Evaluation von Lehrveranstaltungen auch öffentlich zugänglich sein müssen (Paragraph 5 Absatz 3 NHG) und hochschulübergreifende Vergleiche ermöglichen sollen, können sie als Teil eines umfassenden Qualitätssicherungsprozesses verstanden werden. Teil dieses umfassenden Qualitätssicherungsprozesses ist unter anderem auch die Sicherung und Weiterentwicklung der Qualität von Lehre, beispielsweise umgesetzt durch Weiterbildungsmaßnahmen für Lehrende. Darüber hinaus können Kennzahlen der Evaluation von Lehrveranstaltungen auch zur Kontrolle von Zielvereinbarungen und folglich zur Ressourcensteuerung beitragen (Großmann/Wolbring, 2016: 3-4; Rindermann, 2016: 228). Insbesondere der Punkt der Ressourcensteuerung ist im Verständnis des Autors aber eher kritisch zu bewerten, da eine Vielzahl der in dieser Arbeit aufgeführten Fragebögen zur Evaluation von Lehrveranstaltungen methodische Mängel aufwiesen, die zumindest ihre Eignung zur Erfassung von *Lehrqualität* infrage stellen lassen. Wenn aber solche in der methodischen Qualität unbefriedigenden Befunde zur Ressourcensteuerung eingesetzt werden, steigt das Risiko einer Fehlsteuerung.

Gemäß Winteler und Schmolck (1979: 139) kann die Evaluation von Lehrveranstaltungen auf verschiedenen Wegen erfolgen, beispielsweise durch Videoaufzeichnungen und anschließende Beurteilung, Gruppendiskussionen, teilnehmende Beobachtung oder auch durch Gespräche mit Kollegen. Bei großen Teilnehmerzahlen oder einer Vielzahl von zu evaluierenden Lehrveranstaltungen sind die eben genannten Möglichkeiten der Evaluation von Lehrveranstaltungen jedoch kaum praktikabel, weshalb sich ein testökonomischer Fragebogen zur Lösung des Problems anbietet. Fragenbögen zur Lehrevaluation sind im Verlauf der letzten Jahrzehnte zum Standard geworden. In Schmidts und Loßnitzers (2010: 66) empirisch hergeleiteter Definition von Lehrveranstaltungsevaluation – die auch Gegenstand dieser Arbeit ist – spiegelt sich dieser Standard wider: „*Die Lehrveranstaltungsevaluation ist eine spezifische, systematische Form des lehrbezogenen Feedbacks, bei der (1) Studierende (2) schriftlich, d.h. mittels papierhafter Fragebögen oder online (3) in überwiegend standardisierter, d.h. veranstaltungs-, lehrenden- und themenübergreifender Form, (4) anhand eines strukturierten, mehrheitlich geschlossene Items/Fragen umfassenden und um einzelne offene Fragen ergänzten Erhebungsinstruments (5) Einschätzungen zu ausgewählten Aspekten des Verlaufs und der Ergebnisse einer bestimmten*

Lehrveranstaltung oder eines Moduls abgeben." (Schmidt/Loßnitzer, 2010: 66). Auf Basis dieser Arbeitsdefinition wurde nach Fragebögen zur Evaluation von Lehrveranstaltungen recherchiert.

Die Evaluation einzelner Lehrveranstaltungen weist in den Vereinigten Staaten eine längere Tradition auf als in Deutschland (el Hage, 1996: 49; Rindermann, 2016: 228). Bereits Mitte des letzten Jahrhunderts evaluierten rund ein Drittel der amerikanischen Colleges und rund ein Viertel der amerikanischen Universitäten Lehrveranstaltungen (Gustad: 1967: 270-271). In Deutschland fanden Evaluationen von Lehrveranstaltungen erst gegen Ende der 1960er Jahre größere Verbreitung (Webler, 2010: 33). Flächendeckender wurde die Evaluation von Lehrveranstaltungen in Deutschland erst mit Beginn der 1990er-Jahre etabliert (el Hage, 1996: 1; Rindermann, 2016: 228). Trotz der historisch längeren Forschungstradition im amerikanischen Kulturraum, wird im Folgenden nur an ausgewählten Stellen auf englischsprachige Fachliteratur zurückgegriffen. Ein Grund ist die – von Thomas (2001: 220-221) implizit geäußerte Kritik – wonach die kulturübergreifende Generalisierbarkeit empirischer Befunde oftmals ungeprüft unterstellt wird. Die kulturübergreifende Generalisierbarkeit sollte stets empirisch überprüft werden, da eine ungeprüfte Übernahme empirischer Befunde in einen anderen Kulturkreis zu Fehlern führen kann. Häufig haben sich spezifische Regeln, Normen, Verhaltenssysteme und Handlungsvorschriften im Verlauf der Zeit in sozialen Gemeinschaften spezifisch und deshalb unterschiedlich voneinander entwickelt (Thomas: 2016: 16-17). Thomas' Kritik (2001: 220-221) auf Metaebene erscheint zwar berechtigt, führte bei strenger Auslegung aber zu kaum zu interpretierenden Forschungsergebnissen aus anderen Kulturräumen, die zumindest solange kaum zu interpretieren wären, bis deren kulturübergreifende Generalisierbarkeit empirisch überprüft wurde.

Neben dieser generellen Kritik – die bereits den Ausschluss englischsprachiger Literatur an ausgewählten Stellen rechtfertigen würde – nennt Thomas (2016) zusätzlich Beispiele zu Kulturunterschieden, die nach Ansicht des Autors auf die Evaluation von Lehrveranstaltungen übertragen werden können und die folglich als spezifische Kritik an einer ungeprüften, kulturübergreifenden Generalisierung von Befunden zur Evaluation von Lehrveranstaltungen zu verstehen sind. Thomas (2016) schildert eine Vielzahl von Kulturunterschieden, weshalb hier nur zentrale Befunde, die für die Evaluation von Lehrveranstaltungen relevant sind, berichtet werden. In ost- und südostasiatischen Kulturen spielt beispielsweise das „wahren des Gesichts" in zwischenmenschlichen Interaktionen eine besonders wichtige Rolle. Gegen das „wahren des Gesichts" spräche beispielsweise das Stellen von Fragen in Lehrveranstaltungen, bestünde beim Stellen von Fragen doch die Gefahr, Außenstehende könnten Unklarheiten oder unpräzise Fragen als Schwäche auffassen. „Gesicht wahren" umfasst aber nicht nur das Vermeiden gesichtsverlierender eigener Situationen, sondern auch für andere Personen. Nachfragen könnten beispielsweise auch als Schwäche eines Lehrenden verstanden werden, der die Inhalte selbst nicht vollständig beherrscht und deshalb nur ungenau erklärt. In der deutschen Kultur gelten Fragen von Studierenden hingegen als erwünscht, weshalb aktiv am Lehrgeschehen Teilnehmende eher als motiviert wahrgenommen werden als passiv Teilnehmende. Zudem gelten auch kontrovers geführte Debatten im Rahmen einer Lehrveranstaltung im deutschen Kulturraum als eher horizonterweiternd, in anderen Kulturräumen gelten sie mitunter als verletzend (Thomas, 2016: 15-19; 39).

Zurückführen lassen sich die Kulturunterschiede in den Beispielen auf eine unterschiedliche Gewichtung von Sach-, Beziehungs- und Hierarchieorientierung. Weisen Teilnehmer einer Lehrveranstaltung im deutschen Kulturraum durchschnittlich eher eine hohe Sachorientierung auf, weisen Teilnehmer im ost- und südostasiatischen Kulturraum eher eine hohe Beziehungs-

und Hierarchieorientierung auf (Thomas, 2016: 15-19; 39). Im amerikanischen Kulturraum genießt beispielsweise positives Feedback von Fach- und Führungskräften gegenüber ihren Mitarbeitern einen höheren Stellenwert als im deutschen Kulturraum (Thomas, 2016: 54). Da Lehrende eine Art Führungsfunktion im Rahmen einer Lehrveranstaltung ausüben, liegt der Gedanke nahe, dass Lehrende im amerikanischen Kulturraum ebenfalls häufiger – mitunter auch überschwänglicher – loben als Lehrende im deutschen Kulturraum. Aufgrund der genannten Beispiele und aufgrund der generellen Unterschiede zwischen Hochschulsystemen – insbesondere dem deutschen und dem nordamerikanischen (el Hage, 1996: 49) – liegt die Vermutung nahe, dass Kulturunterschiede auch Einfluss auf die Lehr-, Lern- und Evaluationskultur haben, weshalb im Folgenden insbesondere auf den Einbezug von englischsprachigen Fragebögen zur Evaluation von Lehrveranstaltungen verzichtet wird. Englischsprachige Fragebögen stammen einerseits aus einem anderen Kulturraum – was deren konzeptionelle Gestaltung beeinflusst haben dürfte – und andererseits erfordern sie ein methodisch aufwändiges Übersetzungsverfahren, wenn sie in die deutsche Sprache überführt werden sollen. Zudem wird auf empirische Befunde aus anderen Kulturräumen verzichtet, da deren Generalisierbarkeit teilweise unklar ist. Berücksichtigung findet englischsprachige Fachliteratur jedoch hinsichtlich methodischer Stolperfallen, die bei der Konstruktion eines Fragebogens begangen werden können und die unabhängig vom Kulturraum Gültigkeit besitzen.

Im folgenden Abschnitt wird eine Einschätzung anhand der empirischen Befundlage hinsichtlich einer onlinebasierten oder papierbasierten Evaluation von Lehrveranstaltungen vorgenommen. Zudem wird eine Empfehlung hinsichtlich des Zeitraums der Durchführung gegeben.

2.2 Durchführung

Zur Durchführung der Evaluation von Lehrveranstaltungen kommen mehrheitlich zwei Erhebungsmodi zum Einsatz: Eine onlinebasierte oder eine papierbasierte Durchführung. Eine onlinebasierte Durchführung einer Evaluation von Lehrveranstaltungen bietet, verglichen mit einer papierbasierten Durchführung, mehrere Vorteile. Verschiedene Autoren (Pötschke, 2009: 77; Tinsner/Dresel, 2007: 67) nennen als Vorteile beispielsweise eine ökonomischere Datenerfassung und Datenauswertung sowie eine anonymere Erhebungssituation. Auch in der Abteilung Betriebswirtschaft, in der eine papierbasierte Evaluation von Lehrveranstaltungen durchgeführt wird, entsteht durch die Erfassung der ausgefüllten Fragebögen ein erheblicher Verwaltungsaufwand. Zudem ließen sich bei einer onlinebasierten Evaluation von Lehrveranstaltungen beispielsweise keine Rückschlüsse anhand der Handschrift auf die evaluierende Person ziehen und offene Fragen würden ausführlicher beantwortet (Tinsner/Dresel, 2007: 67), wenngleich Meinefeld (2010: 312) widersprüchliche Befunde hinsichtlich der ausführlicheren Beantwortung von Onlinefragebögen berichtet. Der Vorteil der handschriftlichen Anonymität ist plausibel, muss nach Ansicht des Autors aber – auch unter Berücksichtigung des Urteils einer studentischen Expertengruppe – relativiert werden. Die studentische Expertengruppe bestand aus Studierenden der Betriebswirtschaftslehre, die im Sommersemester 2016 das Seminar Personalpsychologie besuchten und mit der Test- und Fragebogenkonstruktion vertraut sind, sowie aus zwei Vertretern des Fachschaftsrats Wirtschaft, die stellvertretend für die Studierenden der Abteilung Betriebswirtschaft sprachen und Bedenken äußerten. Das Vertrauen in die Anonymität von Verhalten im Internet scheint seit dem Bekanntmachen der globalen NSA-Überwachung durch Edward Snowden (Greenwald, 2013, Greenwald/MacAskill/Poitras, 2013) nachhaltig gestört. Die studentische Expertengruppe, die dahingehend in den Prozess der Neugestaltung des Fragebogens eingebunden wurde, als der bestehende Fragebogen hinsichtlich Schwierigkeiten bei der Beantwortung geprüft werden sollte, äußerte vehement Bedenken gegen eine onlinebasierte Durchführung und befürchtete Repressalien bei schlechten Bewertungen von

Lehrenden durch Studierende. Bedenken, die bereits Webler (2005: 68) berichtet und die sozial erwünschtes Antworten begünstigen können.

Tinsner und Dresel (2007: 67) nennen aber nicht nur Vorteile der onlinebasierten Durchführung einer Evaluation von Lehrveranstaltungen sondern auch Nachteile, beispielsweise geringere Rücklaufquoten, mögliche Selektionseffekte, und, daraus resultierend, eine möglicherweise geringere Repräsentativität. Aus der sozialwissenschaftlichen Forschung sind mögliche Probleme bei online erhobenen Daten bekannt, beispielsweise niedrige Rücklaufquoten aufgrund von Motivationsdefiziten, eingeschränkte Repräsentativität aufgrund von Selektionseffekten und reduzierte Datenqualität aufgrund von Antworttendenzen oder fehlenden Antworten (Tinsner/Dresel, 2007: 61). Befunde, die hinsichtlich der Rücklaufquote auch metaanalytisch von Lozar Manfreda et al. (2008: 91-93; n=45) sowie Shih und Fan (2008: 257-259; n=39) nachgewiesen wurden. Im Schnitt ermittelten Lozar Manfreda et al. (2008: 91-93) eine 11 Prozent niedrigere Rücklaufquote von onlinebasierten Datenerhebungen, verglichen mit anderen Erhebungsmethoden wie beispielsweise postalisch basierten Datenerhebungen. Interessanterweise berichten Shih und Fan (2008: 259) widersprüchliche Befunde hinsichtlich der Studien, die Studierende und Fakultätsangehörige einer Hochschule als Testpersonen berücksichtigten. Einzig in der Gruppe der Studierenden und Fakultätsangehörigen erzielten onlinebasierte Datenerhebungen höhere Rücklaufquoten als postalisch basierte Datenerhebungen, wenngleich die durchschnittlich höhere Rücklaufquote von 3 Prozent (n=11) nur geringfügig ausfiel (Shih/Fan, 2008: 257). Inwieweit die metaanalytischen Befunde auf den spezifischen Bereich der Evaluationen von Lehrveranstaltungen übertragbar sind, muss zunächst geklärt werden.

Schroer (2003: 29-30) analysierte anhand einer Längsschnittuntersuchung Unterschiede zwischen papierbasierter und onlinebasierter Evaluation von Lehrveranstaltungen (n=215). Anhand zweier unabhängiger Vier-Felder-Chi-Quadrat-Tests kann Schroer (2003: 48-49) signifikante Unterschiede zwischen papierbasierter und onlinebasierter Evaluation von Lehrveranstaltungen nachweisen. Im Detail konnten in einer Lehrveranstaltung des Grundstudiums (Pearsons X^2= 41,34, p≤.001, ω=.21) und in einer Lehrveranstaltung des Hauptstudiums (Pearsons X^2= 55,75, p≤.001, ω=.25) signifikant geringere Teilnahmequoten bei onlinebasierter Durchführung nachgewiesen werden. Auch Dresels und Tinsners (2008: 193) Befunde sind in Bezug auf onlinebasierte Evaluationen von Lehrveranstaltungen eindeutig, denn betrug die Rücklaufquote der Gesamtstichprobe (n=1.186) in der papierbasierten Durchführung noch 82 Prozent, lag sie in der onlinebasierten Durchführung bei nur 27 Prozent. Eine anschließende inferenzstatistische Analyse auf Veranstaltungsebene (n=7) ergab, dass die onlinebasierten Evaluationen von Lehrveranstaltungen signifikant niedrigere Rücklaufquoten aufwiesen als die papierbasierten (Wilcoxon Z= 2.2, p<0.05). Dresel und Tinsner (2008: 192-206) überprüften zudem, ob weitere Methodeneffekte zwischen den beiden Erhebungsmethoden zutage traten, auf die an dieser Stelle aber nicht näher eingegangen wird, da, abgesehen von den unterschiedlich hohen Rücklaufquoten, sonst nur geringe Effekthöhen vorlagen.

Simon, Zajontz und Reit (2013: 15-16) führten eine Art Vollerhebung durch und analysierten einen Jahrgang, der alle Studiengänge einer Fakultät berücksichtigte. Insgesamt lagen 1.027 auswertbare Evaluationen von Lehrveranstaltungen vor, die aus n=70 Lehrveranstaltungen stammten. Die Auswertung ergab, dass mit der papierbasierten Evaluation von Lehrveranstaltungen eine Rücklaufquote von rund 80 Prozent und mit der onlinebasierten Evaluation von Lehrveranstaltungen eine Rücklaufquote von rund 58 Prozent erreicht wurde. Ein Unterschied, der inferenzstatistisch signifikant ist (t-Wert=4,848, zweiseitig, p≤.001). Tendenziell schlechtere Werte bei deskriptiver Betrachtung erzielen Simonson und Pötschke (2006: 240), Pötschke (2009: 83) sowie Fondel, Lischetzke, Weis und Gollwitzer (2014: 127), die in mehreren onlinebasierten Evaluationen von Lehrveranstaltungen eine Rücklaufquote von 30 bis 48 Prozent

erreichten. Einen prozentualen Vergleichswert zu einer papierbasierten Durchführung bieten die Studien von Simonson und Pötschke (2006), Pötschke (2009) sowie Fondel, Lischetzke, Weis und Gollwitzer (2014) nicht. Meinefeld (2010: 301-303) unterstreicht die bisher dargestellten Befunde. In einer Stichprobe bestehend aus 47 Lehrveranstaltungen und 1.528 auswertbaren Evaluationen von Lehrveranstaltungen, erzielte die onlinebasierte Durchführung deskriptiv eine durchschnittliche Rücklaufquote von 38 Prozent, wohingegen die papierbasierte Durchführung eine durchschnittliche Rücklaufquote von 81 Prozent erreichte (Meinefeld, 2010: 301-303). Ähnliche Werte, die auch Fischer (2014: 42-43) deskriptiv berichtet. Während die onlinebasierte Durchführung Rücklaufquoten zwischen 10 und 51 Prozent hervorrief, erzielte die Evaluation im Anschluss an eine Klausur Rücklaufquoten zwischen 60 und 97 Prozent (Fischer, 2014: 42-43).

Die Ausführungen verdeutlichen, dass die metaanalytisch belegte geringere Rücklaufquote bei onlinebasierter Datenerhebung auch auf Evaluationen von Lehrveranstaltungen übertragen werden kann. Inferenzstatistisch wurden signifikant schlechtere Rücklaufquoten für onlinebasierte Datenerhebungen berichtet und auch deskriptiv lässt sich die Befundlage tendenziell bestätigen. Nun führen verringerte Rücklaufquoten zwar erst bei systematischen Ausfällen zu einer selektiven Stichprobe, nichtsdestotrotz wird empfohlen, die Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft auch zukünftig papierbasiert im Rahmen der Lehrveranstaltungszeit durchzuführen. Für eine Durchführung im Rahmen von Lehrveranstaltungen spricht neben einer vergleichsweise hohen Rücklaufquote auch die Tatsache, dass Lehrveranstaltungen teils schlechte Anwesenheitsquoten aufweisen. Die Evaluation einer Lehrveranstaltung im Grundstudium durch drei Lehrende ergab über mehrere Semester Anwesenheitsquoten von gerade einmal 50 Prozent der planmäßig vorgesehenen Studierenden. Evaluiert man solch eine Lehrveranstaltung mit geringer Teilnahmequote onlinebasiert, kann das weitere Verweigerer – hier in Bezug auf die Evaluation der Lehrveranstaltung – nach sich ziehen, sodass am Ende nur ein kleiner Teil der vorgesehenen Stichprobe die Lehrveranstaltung evaluiert. Zudem äußerte eine studentische Expertengruppe vehement Bedenken und sprach sich gegen eine onlinebasierte Durchführung der Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft aus. Dem Urteil der studentischen Expertengruppe wird ebenfalls Bedeutung zugemessen, da es eine direkte Rückmeldung aus der Zielgruppe der Evaluation von Lehrveranstaltungen darstellt. Nur wenn ein Fragebogen zur Evaluation von Lehrveranstaltungen sowie die Durchführung der Evaluation von der Zielgruppe sozial akzeptiert sind, können zuverlässige Ergebnisse resultieren.

In Summe sprechen somit die empirische Befundlage und die Rückmeldung der studentischen Expertengruppe klar für die papierbasierte Durchführung der Evaluation von Lehrveranstaltungen. Wichtig zu erwähnen ist im Zusammenhang einer papierbasierten Durchführung auch die soziale Akzeptanz bei Lehrenden. Fischer (2014: 42) berichtet etwa für die Lehrenden der Medizinischen Hochschule Hannover, dass die soziale Akzeptanz der papierbasierten Evaluation von Lehrveranstaltungen aufgrund zu lang empfundener Rückmeldezeiten der Ergebnisse zunehmend abgenommen hätte. Auch in der Abteilung Betriebswirtschaft kam es in der Vergangenheit aufgrund organisatorischer Gegebenheiten zu Verzögerungen bei der Rückmeldung der Ergebnisse zur Evaluation von Lehrveranstaltungen an die Lehrenden. Gemäß Paragraph 5 Absatz 1 Satz 4 NHG hat eine Hochschule relevante Aspekte zur internen Evaluation von Lehrveranstaltungen in einer Ordnung festzuhalten. Aktuell ist innerhalb der Ordnung zur internen Lehrevaluation an der Hochschule Hannover kein genauer Zeitpunkt zur Durchführung vorgegeben. Lediglich von einem „*angemessenem Abstand zum Prüfungszeitraum*“, „*spätestens am Ende des Vorlesungszeitraums*“ (Hochschule Hannover, 2006: 2) ist die Rede. Unklar bleibt, wie „angemessen“ definiert ist. Aus Erfahrungen des Autors ist bekannt, dass aktuell

in der Abteilung Betriebswirtschaft zwei Erhebungszeitpunkte zur Auswahl stehen: In der Mitte des Semesters und am Ende des Semesters. Wählt ein Lehrender der Abteilung Betriebswirtschaft als Erhebungszeitpunkt das Ende des Semesters, ist es in der Regel nicht mehr möglich, dass die Ergebnisse noch im laufenden Semester an die Studierenden rückgemeldet werden können. Rindermann (2009: 277) empfiehlt, die Ergebnisse der Evaluation von Lehrveranstaltungen mit den evaluierenden Studierenden zu besprechen, beispielsweise um Ursachen für die Ergebnisse zu ergründen. Auch sollte von Erhebungen am Ende des Semesters abgesehen werden, denn eine bevorstehende Prüfung kann Verzerrungstendenzen begünstigen (Rindermann, 2009: 278). Im Rahmen einer HISBUS-Befragung (Krawietz, 2006), bei der Studierende auch Ratschläge zur Durchführung von Lehrveranstaltungsevaluationen im offenen Antwortformat abgeben konnten, äußerten sich rund 1.200 Studierende. Am häufigsten wünschten sich die Studierenden eine bessere Zugänglichkeit zu den Ergebnissen der Evaluation von Lehrveranstaltungen und nannten auch das gemeinsame Besprechen mit einem Lehrenden als Wunsch (Krawietz, 2006: 22-24). Unklar ist, ob diese Wünsche auch für die Studierenden der Abteilung Betriebswirtschaft gelten.

Da die empirische Befundlage hinsichtlich des idealen Zeitpunkts der Durchführung der Evaluation von Lehrveranstaltungen wenig fundiert ist, kann eine abschließende Empfehlung nur unter Vorbehalt ausgesprochen werden. Die vorliegenden Befunde rechtfertigen am ehesten die Durchführung in der Mitte eines Semesters. Insbesondere Rindermanns (2009) Hinweis auf eine Verzerrungsgefahr „*kurz vor einer Prüfung*“ (Rindermann, 2009: 278) spricht für die Durchführung in der Mitte eines Semesters, wenngleich „kurz“ nicht hinreichend definiert wird und interpretativen Spielraum zulässt. Möchte man sichergehen, den geeignetsten Zeitpunkt für die Durchführung der Evaluation von Lehrveranstaltungen zu finden, empfiehlt sich neben der Berücksichtigung der Empfehlungen, die zusätzliche Erhebung eines Stimmungsbildes der Stakeholder der Abteilung Betriebswirtschaft. Ungeklärt ist derzeit, ob Studierende der Abteilung Betriebswirtschaft eine Rückmeldung der Ergebnisse der Evaluation von Lehrveranstaltungen überhaupt wünschen. Möglicherweise werden auch im Rahmen der Rückmeldung Repressalien bei der schlechten Bewertung eines Lehrenden befürchtet. Aber auch bezüglich der Lehrenden empfiehlt sich die Abfrage eines Stimmungsbilds hinsichtlich des idealen Zeitpunkts der Evaluation von Lehrveranstaltungen. Als Empfehlung wird an dieser Stelle nur formuliert, sich auf einen einzigen Erhebungszeitpunkt zu beschränken. Sollte zukünftig ein Fragebogen zur Evaluation von Lehrveranstaltungen eingesetzt werden, der zuverlässige Messungen ermöglicht, könnten beispielsweise auch interindividuelle Vergleiche durchgeführt werden. Bei unterschiedlichen Erhebungszeitpunkten wäre dies nicht möglich.

Im folgenden Abschnitt werden Variablen diskutiert, die lehrunabhängige Einflüsse ausüben und folglich verzerrend auf die Evaluation von Lehrveranstaltungen wirken können.

2.3 Verzerrungsvariablen

Es gibt eine Vielzahl von Variablen, die verzerrend auf die Evaluation von Lehrveranstaltungen wirken können. Von einer Verzerrungsvariablen wird gesprochen, wenn eine Variable das Urteil der Evaluation von Lehrveranstaltungen beeinflusst, ohne dass sie in einem inhaltlichen Zusammenhang mit dem erhobenen Konstrukt *Lehrqualität* steht, beispielsweise die *Attraktivität* eines Lehrenden. El Hage (1996: 51-84), Koch (2004: 305), Rindermann (2009: 181) und Wolbring (2013: 117-122) nennen beispielhaft mögliche Verzerrungsvariablen, die die Validität von Lehrveranstaltungsevaluationen infrage stellen können: der *Besuchsgrund* eines Studierenden, *Interesse* am Thema, das *Schwierigkeitsniveau* einer Lehrveranstaltung, der *Arbeitsaufwand* zum Bestehen, das *Geschlecht* eines Lehrenden oder der Studierenden, das *Leistungsni-*

veau der Teilnehmer, die *Sympathie* für einen Lehrenden und so weiter. Alle genannten Variablen stehen mit der *Lehrqualität* in keinem inhaltlichen Zusammenhang, weshalb sich ihre Berücksichtigung empfehlen kann, insbesondere dann, wenn weitreichende Entscheidungen auf Grundlage der Evaluation von Lehrveranstaltungen getroffen werden. Rindermann (2009: 183-185) hat eine Übersicht zu zentralen Befunden der Forschung zu Verzerrungsvariablen im Bereich der Evaluation von Lehrveranstaltungen erstellt, wenngleich er darauf verweist, keine vollständige Übersicht zu liefern. Eine vollständige Übersicht würde auch den Rahmen dieser Arbeit übersteigen. Vielmehr soll nachfolgend skizziert werden, welche möglichen Verzerrungsvariablen am ehesten berücksichtigt werden könnten.

Rindermann (2009: 185) fasst den internationalen Forschungsstand wie folgt qualitativ zusammen: Personenbezogene Variablen wie *Alter*, *Geschlecht* und *Semesterzahl* üben keine verzerrenden Einflüsse aus, unabhängig davon, ob sich personenbezogene Variablen auf einen Lehrenden oder Studierende beziehen. Vernachlässigbare Einflüsse üben beispielsweise das *Leistungsniveau* von Studierenden, *Persönlichkeitsmerkmale* und die *wahrgenommene* sowie *tatsächliche Ähnlichkeit* zwischen einem Lehrenden und Studierenden aus (Rindermann, 2009: 185). Substanzielle Einflüsse üben – unabhängig von der kausalen Richtung – das *Vorinteresse* am Thema, das *Thema* selbst und der *Besuchsgrund* einer Lehrveranstaltung aus (Rindermann, 2009: 185). Einflüsse, die, möchte man die Evaluation von Lehrveranstaltungen als Entscheidungsgrundlage verwenden, berücksichtigt werden sollten. Beispielsweise sollten keine Vergleiche zwischen Wahlpflichtmodulen und Pflichtmodulen angestellt werden, denn sofern mögliche Verzerrungsvariablen nicht rechnerisch berücksichtigt und ihr Einfluss korrigiert wurde, ist unklar, ob Differenzen aufgrund von Unterschieden in der *Lehrqualität* auftreten oder aufgrund von Verzerrungsvariablen. Weil eine kulturübergreifende Betrachtung von Verzerrungsvariablen (el Hage, 1996: 51-67; Rindermann, 2009: 189-194; Wolbring, 2013: 120-123) problematisch ist, werden im Folgenden ausgewählte Befunde zu Verzerrungsvariablen präsentiert, die innerhalb des deutschen Kulturraums entstanden sind. Die Ausführungen konzentrieren sich auf Rindermann (2009), der a) innerhalb des deutschen Kulturraumes einschlägig ist und b) valide Fragebögen für seine Analysen verwendet hat.

Rindermanns (2009: 188-201) Analysen zu Verzerrungsvariablen zeigen substanzielle Zusammenhänge in Bezug auf das *Thema* einer Lehrveranstaltung. Waren Studierende thematisch interessiert, beurteilten sie das *Lehrverhalten* (Rohwert: N=24.400, r=.35; Veranstaltungsmittel: N=999, r=.43) und die *Lehreffektivität* (Rohwert: N=24.400, r=.57; Veranstaltungsmittel: N=999, r=.65) positiver. Befunde, die unter Verwendung des *HILVE-II*-Fragebogens (Rindermann, 2009: 388-391) nur teilweise repliziert wurden (Rindermann, 2009: 193-194). Das zeigt die Problematik in diesem Forschungsbereich auf, weil sich Befunde zu Verzerrungsvariablen selbst unter Verwendung der Weiterentwicklung desselben Fragebogens vom *HILVE* (Rindermann/Amelang, 1994) hin zum *HILVE-II* (Rindermann, 2009: 388-391) nur teilweise replizieren lassen. Und das, obwohl beide Fragebögen dieselbe Grundlage aufweisen und im gleichen Kulturraum verwendet wurden. Veranschaulicht man sich nun (vergleiche Tabelle 1), dass mit sehr vielen Fragebögen Analysen hinsichtlich Verzerrungsvariablen durchgeführt wurden, wird offensichtlich, weshalb die Forschung zu Verzerrungsvariablen als „*diffizil*“ (Rindermann, 2009: 187) beschrieben werden kann.

Nichtsdestotrotz lassen sich Befunde der Forschung zu Verzerrungsvariablen sinnvoll aufbereiten. Möchte man mögliche Verzerrungsvariablen auf einen zu berücksichtigenden Kern reduzieren, empfiehlt sich die Betrachtung der Variablen *Vorinteresse*, *Besuchsgrund* und *Anforderungsniveau*. Möchte man aber beispielsweise einen besonders ökonomischen Fragebogen trotz möglicher Informationsverluste verwenden, lässt sich auch auf die Erhebung von

möglichen Verzerrungsvariablen verzichten, denn summa summarum ist der Einfluss möglicher Verzerrungsvariablen oftmals geringer als befürchtet (Rindermann, 2009: 201; Wolbring, 2013: 120-121). Zusammengefasst lässt sich festhalten:

- Ist a) ein möglichst hoher Informationsgewinn das Ziel, sollten auch Verzerrungsvariablen bedacht und innerhalb eines Fragebogens berücksichtigt werden.
- Ist b) ein besonders ökonomischer Fragebogen das Ziel, kann auch auf die Berücksichtigung möglicher Verzerrungsvariablen verzichtet werden, ohne dass der Informationsgewinn gen Null tendierte.

Je nach Zielsetzung können auch Positionen zwischen den beiden Extremen vertreten werden. Im Rahmen dieser Arbeit ist – wie in *3 Methodik* noch zu sehen sein wird – ein besonders ökonomischer Fragebogen ein Ziel.

Im folgenden Abschnitt werden ausgewählte Fragebögen zur Evaluation von Lehrveranstaltungen vorgestellt, die hinsichtlich ihrer Eignung geprüft wurden.

2.4 Fragebögen

Im deutschen Sprachraum gibt es mehrere Fragebögen zur Evaluation von Lehrveranstaltungen. Gelegentlich erschien es bei den Recherchen auf Hochschulwebseiten als habe jede Hochschule ihren eigenen Fragebogen zur Evaluation von Lehrveranstaltungen konzipiert. Hansen, Hennig-Thurau und Wochnowski (1997: 377) zählten bereits in den 1990er Jahren rund 170 Fragebögen, die allein in den Wirtschaftswissenschaften an Hochschulen verwendet wurden. Vergleiche auch Reissert (1992) für eine Zusammenstellung überwiegend ad-hoc erstellter Fragebögen. Um eine systematische Recherche nach Fragebögen zur Evaluation von Lehrveranstaltungen durchzuführen, bei denen psychometrische Gütekriterien bei der Konstruktion berücksichtigt wurden, fand die Recherche nicht auf den Webseiten einzelner Hochschulen statt, sondern in ausgewählten Datenbanken und einschlägigen Publikationen. Wichtig ist, dass mit dem gewählten Vorgehen keine Herabwürdigung oft sorgfältig erstellter, unveröffentlichter Fragebögen verbunden ist. Die Recherche in Datenbanken und einschlägigen Publikationen ist schlicht der kaum zu überschaubaren Anzahl unveröffentlichter Fragebögen zur Evaluation von Lehrveranstaltungen geschuldet. Einschlägige Publikationen erwiesen sich zudem als wertvolle Informationsquelle, weil andere Autoren (Braun/Gusy, 2008: 154-155; Diehl, 2003: 29; Rindermann, 2009: 83-85; Schmidt/Loßnitzer, 2010: 61; Westermann/Spies/Heise/Wollburg-Claar, 1998: 142) mitunter die Ergebnisse ihrer Recherchen nach Fragebögen übersichtlich veranschaulicht haben. Recherchiert wurde in ausgewählten Datenbanken mit den zwei Suchbegriffen *Evaluation* und *Lehrevaluation*. Der Suchbegriff *Evaluation* wurde in der Datenbank *Fachportal Pädagogik* verwendet. Mit dem engeren Suchbegriff *Lehrevaluation* wurde recherchiert, wenn sich die Anzahl der Treffer auch bei intelligenter Suche nicht auf unter 1.000 Treffer senken ließ. Recherchiert wurde mit dem Suchbegriff *Lehrevaluation* in den Datenbanken *PSYNDEX*, *OLC Hochschulwesen*, *IBZ Online*, *WISO*, *EconBiz*, *Scopus* und *Web of Science*. Das Resultat der Recherchen ist in Tabelle 1 zusammengefasst. Erfasst wurden Fragebögen, die a) anhand der Datenbankrecherche identifiziert wurden – unabhängig davon, ob der jeweilige Fragebogen mitveröffentlicht wurde oder nur psychometrische Gütekriterien veröffentlicht wurden, anhand derer anschließend der jeweilige Fragebogen recherchiert werden konnte – oder b) in einschlägiger Literatur erwähnt wurden. Aufgeführt sind in Tabelle 1 die Fragebögen, deren jeweilige Antwortkategorien und die Anzahl der Items. Da die Anzahl der Items von Autoren mitunter unterschiedlich berichtet wird – einige Autoren zählen Fragen mit soziodemographischem Inhalt oder offenem Antwortformat nicht mit, andere hingegen tun es – wurde die Anzahl an Items selbst gezählt. Gezählt wurden nur Fragen mit geschlossenem Antwortfor-

mat, die sich auf die Evaluation einer Lehrveranstaltung beziehen. Nicht gezählt wurden beispielsweise Fragen mit soziodemographischem Inhalt oder offenem Antwortformat. Will man also beispielsweise die Länge des neu erstellten Fragebogens (vergleiche *4 Ergebnis*) mit den aufgeführten Fragebögen vergleichen, sollte man gedanklich die fehlenden soziodemographischen Fragen und Fragen mit offenem Antwortformat berücksichtigen. Zudem sind die dimensionale oder faktorielle Struktur der jeweiligen Fragebögen und psychometrische Maße der internen Konsistenz mit in Tabelle 1 aufgeführt.

Nicht zu allen Fragebögen berichteten die jeweiligen Autoren Maße der internen Konsistenz, weshalb bei einem Teil der Fragebögen keine Werte aufgeführt sind. Weitere Reliabilitätsmaße wie beispielsweise die Retestreliabilität sind nicht aufgeführt. Reliabilitätsmaße werden von den aufgeführten Autoren seltener berichtet als das häufiger aufgeführte Maß Cronbach α, weshalb in Tabelle 1 das psychometrische Gütekriterium aufgeführt wird, das am häufigsten berichtet wurde. Ebenfalls nicht mit in Tabelle 1 aufgeführt sind die Ergebnisse faktorenanalytischer Überprüfungen der Fragebögen. Der Grund dafür, dass faktorenanalytische Überprüfungen nicht mit aufgeführt sind, ist, neben dem selteneren Ausweis, die aus statistischer Sicht nicht gegebene Vergleichbarkeit einzelner faktorieller Berechnungen. Wurden in der Vergangenheit oftmals einfache konfirmatorische Faktorenanalysen gerechnet, werden mittlerweile eher multilevel konfirmatorische Faktorenanalysen eingesetzt (Sengewald/Vetterlein, 2015: 120-121; Ziegler/Weis, 2015: 114).

Nicht mit aufgeführt in Tabelle 1 sind zudem Fragebögen, die zwar im Rahmen der Evaluation von Lehrveranstaltungen verwendet werden, die aber den Kompetenzerwerb der Studierenden operationalisieren. Messungen des Kompetenzerwerbs sind unter historischer Betrachtung ein eher junges Themenfeld der Evaluation von Lehrveranstaltungen und sollen das Ergebnis von Lehr- und Lernprozessen quantifizieren (Pohlenz/Oppermann, 2010: 4). Im Rahmen des Bologna-Prozesses wurde unter anderem der berufsqualifizierende Aspekt von Bachelorstudiengängen (Europäische Bildungsminister, 1999: 3) festgelegt, den Braun (2007: 74) als Kompetenzvermittlungspflicht interpretiert. Fragebögen zum Kompetenzerwerb (vergleiche Braun/Gusy/Leidner/Hannover, 2008: 34-35; Paechter/Maier/Macher, 2011: 6) stellen eine eigene Klasse an Fragebögen zur Evaluation von Lehrveranstaltungen dar, die lediglich das Ergebnis des Lehr- und Lernprozesses betrachten, und zudem umfangreich sind. Auch erscheinen Fragebögen zum Kompetenzerwerb deshalb ungeeignet, weil in allen Lehrveranstaltungen ein und derselbe Fragebogen eingesetzt werden soll. Aber nicht alle Lehrveranstaltungen zielen auf den Erwerb derselben Kompetenzen ab, weshalb verschiedene Fragebögen zur ökonomischen Kompetenzmessung eingesetzt werden müssten. Es erscheint daher sinnvoller, Kompetenzmessungen auf die Lehrveranstaltungen zu beschränken, die den Erwerb spezifischer Kompetenzen explizit zum Ziel haben, beispielsweise Lehrveranstaltungen zu sozialer Kompetenz. Wenngleich an dieser Stelle ausdrücklich auch auf einen kritischen Aspekt von Kompetenzmessungen generell hingewiesen werden soll. Schuler (2014: 85) nimmt aus personalpsychologischer Sicht Stellung zum Kompetenzbegriff und warnt ausdrücklich davor, „*ein Sammelsurium von Fähigkeiten, Fertigkeiten und Erfahrungen, Verhaltensbereitschaften und Verhaltensergebnissen zusammenzustellen, dabei Voraussetzungen und Konsequenzen – also Prädiktoren und Kriterien – zu vermischen, taxonomische Oberbegriffe und ihre Teilaspekte aneinanderzureihen, Synonyme nicht als solche zu erkennen und auch in jeder weiteren Hinsicht hinter alle Entwicklungen zurückzufallen, die die Arbeits- und Anforderungsanalyse in den letzten 50 Jahren genommen hat.*“ (Schuler, 2014: 85). Schulers (2014: 85) Worte sind unmissverständlich gewählt, sollen im Verständnis des Autors aber nur teilweise auf Kompetenzmessungen im Rahmen der Evaluation von Lehrveranstaltungen übertragen werden, die lediglich eine grobe Bestimmung zum Ziel haben und eine Rückmeldung an Lehrende und Studierende ermöglichen sollen. Im

Rahmen einer Arbeits- und Anforderungsanalyse im Berufsleben können jedoch auch Konsequenzen für Mitarbeiter die Folge sein, weshalb Schuler (2014: 85) aus personalpsychologischer Sicht berechtigt das methodisch nicht korrekte Vorgehen moniert. Für die Konzeption eines Fragebogens zur Kompetenzmessung im Rahmen der Evaluation von Lehrveranstaltungen sollten Schulers (2014: 85) Ausführungen aber zumindest bedacht werden. Auch in Hinblick darauf, dem akademischen Nachwuchs keine methodisch unsauberen Messungen vorzuleben, die später bedauernswerterweise nur noch hartnäckiger von eben jenem Nachwuchs in der beruflichen Praxis verankert werden.

Insbesondere vor dem Hintergrund der Neugestaltung der Lehrevaluation durch das Zentrum für Studium und Weiterbildung der Hochschule Hannover sei der Aspekt methodisch nicht überzeugender Messungen erwähnt. Von diesem Zentrum für Studium und Weiterbildung wird geplant, die klassische Lehrevaluation durch eine outputorientierte Evaluation abzulösen, die unter anderem Kompetenzmessungen beinhaltet. Dazu soll nach aktuellem Stand einerseits der methodisch schlecht konzipierte alte Fragebogen verkürzt werden (Seite 1) und andererseits eine Selbsteinschätzung des Kompetenzfortschritts durch Studierende (Seite 2) vorgenommen werden, die anschließend mit den Einschätzungen des Kompetenzfortschritts durch Lehrende abgeglichen werden sollen (ZSW Hochschuldidaktik: ohne Jahr: 1-3). Nach Ansicht des Autors ist dieses Vorgehen – insbesondere bei der gleichzeitigen Durchführung beider Messungen – mit einigen methodischen Stolperfallen verbunden, die am Beispiel des Durchführungszeitpunkts beschrieben werden. Führte man die neu geplante Lehrevaluation zur Mitte des Semesters durch, bestünde womöglich die Problematik, die Messung mit einem erheblichen Messfehler zu versehen. Studierende hätten ihren Kompetenzfortschritt bereits nach der Hälfte der absolvierten Lehrinhalte zu beurteilen, obwohl sich vermutlich ein Teil der Studierenden erst gegen Ende des Semesters intensiver mit den gelehrten Inhalten beschäftigt. Auch hätten Lehrende den Kompetenzfortschritt von Studierenden nach der Hälfte der gelehrten Inhalte zu beurteilen, obwohl die Lehrenden ihre Veranstaltung und die damit verbundene Zielsetzung in der Regel für ein vollständiges Semester konzipiert haben, was die Evaluation erschweren dürfte. Führte man die neu geplante Lehrevaluation hingegen am Ende des Semesters durch, bestünde unter anderem die Problematik, eine Verzerrungsgefahr im Fragebogen zur Evaluation der Lehrveranstaltung hervorzurufen (Rindermann, 2009: 278), wie unter *2.2 Durchführung* bereits dargelegt. Nach Benotung der Studierenden wiederum müsste die neugestaltete Evaluation onlinebasiert durchgeführt werden, was ebenfalls methodisch heikel ist, wie oben erläutert wurde. Fragwürdig bleibt, weshalb subjektive Einschätzungen herangezogen werden sollen, wenn objektivere outputorientierte Messungen, wie beispielsweise Leistungsnachweise, vorliegen. Es besteht die Gefahr, eine lang bewährte Klasse an Fragebögen mit methodisch zweifelhafter Kompetenzmessung zu verwässern. Das Argument der Vermeidung des Abstrafens von Lehrveranstaltungen und Lehrenden wird nicht geteilt (ZSW Hochschuldidaktik, ohne Jahr: 2), denn bereits seit Längerem werden die Lehrveranstaltungen an der Hochschule Hannover – insbesondere der Abteilung Betriebswirtschaft – ausgesprochen gut evaluiert. Im Wintersemester 2014/2015 sowie dem Sommersemester 2015 wurden folgende Werte in der Abteilung Betriebswirtschaft erzielt: Inhalt der Lehrveranstaltung=1,8; Bewertung der Lehrenden=1,6 und Didaktik der Lehrveranstaltung=2,0. Den erzielten Werten liegt ein Schulnotenprinzip von 1 (maximale Veranstaltungsgüte) bis 6 (minimale Veranstaltungsgüte) zugrunde (Litzcke, 2015). Im Wintersemester 2015/2016 und dem Sommersemester 2016 wurden vergleichbare Werte erzielt. Zudem sollte die Kritik erlaubt sein, eine schlechte Lehrveranstaltung auch entsprechend zu evaluieren, denn nur so lassen sich schließlich Missstände aufdecken. Auch sei aus kollegialer Sicht empfohlen, methodisch zweifelhafte Formulierungen von Items zu vermeiden. Im Musterbeispiel des *Lehrendenfragebogens* (ZSW Hochschuldidaktik: ohne

Jahr: 2) sind Items aufgeführt, die nur schwer für Lehrende zu beurteilen sind. Insbesondere in gut besuchten Lehrveranstaltungen des Grundstudiums sind Items wie „Die Studierenden bewältigen fachliche Probleme wesentlich besser“ und „Informationen beschaffen sich die Studierenden vermehrt selbstständig“ für Lehrende kaum verlässlich zu beurteilen. Was, wenn sich nur ein Teil der Studierenden aktiv beteiligt? Es könnte einem Ratespiel gleichkommen als Lehrender die Entwicklung der fachlichen Kompetenz von einem Teil der Studierenden auf die gesamte Gruppe zu extrapolieren. Auch kann nur schwer beurteilt werden, ob sich Studierende vermehrt selbstständig Informationen beschaffen. Mitunter reicht bereits ein aktiver Studierender, der Informationen beschafft und in sozialen Netzwerken für alle bereitstellt, um Selbstständigkeit der gesamten Gruppe vorzutäuschen. Anhand eines weiteren Items sollen Lehrende beurteilen, ob das Interesse der Studierenden an fachlichen Problemen deutlich zugenommen hat. Ungeklärt bleibt bei Fragen dieser Art, wie „deutlich“ zu interpretieren ist. Auch ist unklar, was ein Lehrender auswählen soll, wenn Studierende auf einem gleichbleibend sehr hohen Niveau Interesse an fachlichen Problemen zeigen. Strenggenommen müsste ein Lehrender das Item als völlig unzutreffend beurteilen, was absurderweise wohl als schlechter Output der Lehrveranstaltung zu werten wäre. Zusammengefasst soll keine generelle Ablehnung gegen eine outputorientierte Messung ausgesprochen werden, sondern lediglich der Hinweis gegeben werden, eine neue Form der Lehrevaluation auf methodisch festem Fundament aufzubauen.

In Tabelle 1 sind neben den Namen der Fragebögen auch Informationen zum Antwortformat, zur Anzahl der Items und zur dimensionalen oder faktoriellen Struktur der Fragebögen mit aufgeführt. In der letzten Spalte sind Maße der internen Konsistenz ausgewiesen, sofern die Autoren davon berichteten. Es sind jedoch nicht alle Fragebögen mit aufgeführt, die im Rahmen der Recherche identifiziert wurden. Nicht mit aufgeführt sind beispielsweise Tröster, Gundlach und Moschner (1997: 112-113), die zwar im Rahmen einer Befragung von Studierenden zu Diplomarbeiten auch eine Skala zur Beurteilung von Lehrenden berücksichtigen, deren 16 Items aber leider nur teilweise veröffentlicht haben. Da Trösters, Gundlachs und Moschners (1997) Fragebogen zur Beurteilung mehrerer Lehrender und nicht nur eines einzelnen Lehrenden konzipiert wurde, wurde davon abgesehen den Fragebogen zu beschaffen. Ebenfalls nicht mit aufgeführt ist der *Fragebogen zur Wärmeübertragung* (Alvensleben/Morsch/Schirmer, 1978: 153-157), der speziell für den naturwissenschaftlichen Bereich konzipiert ist. Zwar wurden generell auch Fragebögen anderer Fachdisziplinen berücksichtigt, der *Fragebogen zur Wärmeübertragung* (Alvensleben/Morsch/Schirmer, 1978: 153-157) ist aber insofern ungeeignet, da a) Abstraktionsbereiche operationalisiert werden – die speziell für technische Lehrveranstaltungen gelten und die zunächst einmal für Lehrveranstaltungen der wirtschaftswissenschaftlichen Disziplinen widerspruchsfrei identifiziert werden müssten – darüber hinaus umfassen b) Teile des Fragebogens nicht die *Lehrqualität* im eigentlichen Sinne, sondern operationalisieren Wunschvorstellungen wie Lehre durchgeführt werden sollte. Auch Reisserts (1992) umfangreiche Zusammenstellung verwendeter Fragebögen blieb unberücksichtigt, da die Fragebögen überwiegend ad-hoc erstellt wurden und keine testtheoretischen Überprüfungen umfassen.

Tabelle 1: Fragebögen zur Evaluation von Lehrveranstaltungen – Reihenfolge nach Veröffentlichungsdatum (eigene Darstellung).

	Fragebogen	*Antwortformat*	*Anzahl Fragen*	*Dimensionen/Faktorenstruktur*	*interne Konsistenz*
1	Fragebogen für Vorlesungen (LVV) (Müller-Wolf, 1977: 197-203)	sechsstufig, -3 (starke Ausprägung) bis +3 (starke Ausprägung), dimensionale Darstellung von Gegensatzpaaren	53	-	-
2	Langform VB-Psych (Diehl/Kohr, 1977: 66-70) Kurzform VB-VOR (Diehl, 2002: 6-7)	Langform: vierstufig, vollständig beschriftet, von *stimmt* bis *stimmt nicht* Kurzform: vierstufig, vollständig beschriftet, von *stimmt nicht* bis *stimmt*	40 Langform; 16 Kurzform	1) Relevanz und Nützlichkeit der Veranstaltungsinhalt, 2) Verhalten des Dozenten gegenüber den Veranstaltungsteilnehmern, 3) Angemessenheit von Schwierigkeit und Umfang der Veranstaltungsinhalte, 4) Methodik und Aufbau der Veranstaltung	Langform: - Cronbachs α der Faktoren von .91 bis .95 (Diehl/Kohr, 1977: 69) - Cronbachs α der Faktoren von .84 und .92 (Kleine/Merkens, 1979: 150) - Cronbachs α der Faktoren von .74 bis .90 (Hofmann, 1990: 50) Kurzform: - Cronbachs α der Faktoren von .83 bis .87 (Diehl, 2002: 4-5)
3	Fragebogen zur gemeinsamen Unterrichtskritik (Rieck, 1978: 221-224)	siebenstufig, endpunktskaliert, von *trifft vollkommen zu* bis *trifft überhaupt nicht zu*	7	-	-
4	Fragebogen zur Unterrichtssituation (Sommer, 1978: 174; 183)	unterschiedliches Antwortformat für die Skalen Ist-Zustand, Soll-Zustand und Relevanz	30 je Skala (Ist, Soll und Wichtigkeit), insgesamt 90	1) praktische Relevanz, 2) Dozent auf Überlegenheit bedacht, 3) effiziente Stoffvermittlung, 4) Eigenaktivität der Studenten, 5) emotionale Atmosphäre Dozent – Student (Sommer, 1978: 177)	-

5	Fragebogen zur Lehrveranstaltungsbeurteilung (Winteler/ Schmolck, 1979: 142-144)	fünfstufig, vollständig beschriftet, von *trifft voll zu* bis *trifft überhaupt nicht zu*	61	1) Fachdiskussion, 2) Wiederholungen, 3) Stoffauswahl und Stoffgliederung, 4) Klima, 5) Schwierigkeit, 6) Relevanz, 7) Leerlauf, 8) Stoffverständnis, 9) Aktivierung, 10) Lehrziele, 11) Akustische Verständlichkeit, 12) Beispiele, 13) Rückmeldung Lernfortschritt (Winteler/Schmolck, 1979: 150).	-
6	Kursbeurteilungsbogen (KBB) (Reischmann, 1995)	fünfstufig, vollständig beschriftet, von *trifft ganz und gar zu* bis *trifft gar nicht zu*	39	1) Stoffbeherrschung, 2) Förderung der Lernbereitschaft, 3) Lernunterstützung, 4) Klima (Reischmann, 1995: 274)	- Cronbachs α der Faktoren von .76 bis .87 (Reischmann, 1995: 275)
7	Lehrverhaltensinventar (Astleitner/Krumm, 1996: 18-20)	fünfstufig, vollständig beschriftet, von *(fast) nie* bis *(fast) immer*	57 Langform; 25 Kurzform	1) Sprache, 2) Nonverbales Verhalten, 3) Erklärung des Lehrstoffes, 4) Organisation, 5) Motivierung, 6) Aufgabenorientierung, 7) Belohnung, 8) Partizipation (Astleitner/Krumm, 1996: 15)	- Langform Cronbachs α = .90 - Kurzform Cronbach α = .84 - Kurzform Cronbach α = .80 (Astleitner/Krumm, 1996: 15)
8	TEACH-Q (Hansen et al., 1997: 389-392)	fünfstufig, endpunktskaliert, von *trifft vollkommen zu* bis *trifft überhaupt nicht zu*	68	1) Nachvollziehbarkeit der Veranstaltungsinhalte, 2) Auftreten und Wirken des Dozenten, 3) Vorlesungsinhalte, 4) Praxisbezug der Vorlesung, 5) Beteiligungsmöglichkeiten der Studenten, 6) Prüfungsvorbereitung, 7) Arbeitsbedingungen, 8) Vortragsstil, 9) Studentisches Verhalten während der Veranstaltung (Hansen et al., 1997: 383)	- Cronbachs α der Faktoren von .57 bis .91 (Hansen et al., 1997: 381)

9	Fragebogen zur Evaluation von Lehrveranstaltungen (Fiedler/Billmann-Mahecha, 1997: 7-9)	fünfstufig, endpunktskaliert, von *trifft völlig zu* bis *trifft gar nicht zu*	24	1) Organisatorische und didaktische Kompetenz, 2) Bereitschaft zur Kooperation und Akzeptanz	- Cronbachs α der Faktoren von .87 bis .90
10	Fragebogen zur Evaluation von Lehrveranstaltungen (Lohnert/Rolfes, 1997)	fünfstufig, vollständig mit Symbolen beschriftet, von ++ bis --, mit einer sechsten Möglichkeit *kann ich nicht beurteilen*	27	-	-
11	Fragebogen zur Evaluation des Lehr- und Lernverhaltens (FELL-V) (Moosbrugger et al., 1997: 10-13)	vierstufig, vollständig beschriftet, von *trifft nicht zu* bis *trifft zu*, mit zwei weiteren Möglichkeiten *nicht sinnvoll* und *nicht wichtig*	25	1) didaktische Kompetenz des Veranstalters, 2) Engagement/Motivation der Studierenden, 3) Förderung eigenständiger Mitarbeit/Motivation der Studierenden, 4) individuelle Betreuung der Studierenden, 5) Inhalte der Veranstaltung, 6) Lernerfolg der Studierenden, 7) Qualität von Referaten, 8) Räumlichkeiten und technische Ausstattung, 9) Sicherung grundlegender Studientechniken	-
12	Fragebogen zur Beurteilung einer Lehrveranstaltung (FB-LV) (Westermann et al., 1998: 149-151)	Antwortkontinuum von *trifft überhaupt nicht zu* bis *trifft vollständig zu* (Westermann et al., 1998: 138)	55	1) allgemeine Zufriedenheit mit einer Lehrveranstaltung, 2) Zufriedenheit mit den Inhalten, 3) Zufriedenheit mit den Studienbedingungen, 4) Bewältigung der Studienbelastungen (Westermann et al., 1998: 138)	- gemäß den Autoren (Westermann et al., 1998: 159) eine ausreichende Reliabilität
13	Kommunikations-Instrument für die Evaluation von Lehrveranstaltungen (KIEL) (Gediga et al., 2000: 84-87)	fünfstufig, vollständig beschriftet, von *trifft völlig zu* bis *trifft gar nicht zu*, mit einer sechsten Möglichkeit *kann ich nicht beurteilen*	14	1) Verständlichkeit, 2) Struktur/Gliederung, 3) Materialien, 4) Interesse (Gediga et al., 2000: 84-88)	- Cronbachs α der Faktoren von .68 bis .83 (Gediga et al., 2000: 74)

14	Fragebogen zur Evaluation von Vorlesungen (FEVOR) (Staufenbiel, 2000: 178)	fünfstufig, mit variierender Bezeichnung des Antwortformats	20	1) Planung und Darstellung, 2) Umgang mit Studierenden, 3) Interessantheit und Relevanz, 4) Schwierigkeit und Umfang	- Cronbachs α der Faktoren auf Personenebene = von .67 bis .77 - Cronbachs α der Faktoren auf Veranstaltungsebene von .81 bis .86 (Staufenbiel, 2000: 174)
15	Fragebogen zur globalen Lehrveranstaltungsevaluation (Wolf et al., 2001: 102-104)	fünfstufig, endpunktskaliert, von *trifft völlig zu* bis *trifft nicht zu* (Wolf et al., 2001: 99)	19	1) Variablen zur LV-Identifikation, 2) fachliche und didaktische Kompetenz, 3) Engagement des LV-Leiters, 4) Einsatz von Lehr- und Lernbehelfen, 5) Struktur und Organisation der LV, 6) Umfeld und studierendenbezogene Variablen, 7) Gleichbehandlung, 8) Gesamtbeurteilung	-
16	Feedback zur Vorlesung (FzV) (Souvignier/Gold, 2003: 134-135)	vierstufig, endpunktskaliert, von *stimmt* bis *stimmt nicht*	18	1) Struktur, 2) Interesse, 3) Aufmerksamkeitssteuerung, 4) Wiederholen/Üben, 5) Hinweise zum Lernen (Souvignier/Gold, 2003: 137-138)	-
17	Trierer Inventar zur Lehrevaluation (TRIL) (Gläßer et al., 2002; Gollwitzer/Schlotz, 2003)	sechsstufig, endpunktskaliert, von *trifft überhaupt nicht zu* bis *trifft voll und ganz zu*	34; adaptierte Version = 16 Items (Fondel et al., 2014: 127)	1) Struktur und Didaktik, 2) Anregung und Motivation, 3) Interaktion und Kommunikation, 4) Persönlicher Gewinn durch die Veranstaltung, 5) Anwendungsbezug, 6) Gesamtbeurteilung (Gollwitzer et al., 2006: 92; Gollwitzer/Schlotz, 2003: 120-121).	- Cronbachs α der Faktoren von .73 bis .89 (Gollwitzer/Schlotz, 2003: 122).
18	Fragebogen zur Lehrevaluation in der Medizin (Berger et al., 2003: 73)	sechsstufig, endpunktskaliert, von *stimmt* genau bis *stimmt nicht*	13	1) Didaktik (Berger et al., 2003: 77)	- Cronbachs α des Faktors Didaktik = .93 (Berger et al., 2003: 77)

19	Fragebogen zur Erfassung des Dozierendenverhaltens (Koch, 2004: 399-411)	überwiegend siebenstufig, endpunktskaliert, von *sehr wenig* bis *sehr stark*	63 Sollwert-Version; 90 Istwert-Version;	1) Verhältnis, 2) Diskussion, 3) Flexibilität, 4) Lenkung, 5) Verständlichkeit, 6) Organisation, 7) Elaboration, 8) kritisches Prüfen, 9) Wiederholung, 10) Planung, 11) Überprüfung, 12) Regulierung, 13) Stimulierung, 14) Autonomie, 15) Kompetenzunterstützung, 16) didaktisch-konstruktivistische Lehraspekte, 17) Aufmerksamkeit wecken, 18) Variabilität, 19) Enthusiasmus, 20) Referatsbetreuung	- Cronbachs α der Faktoren von .60 bis .98 (Koch, 2004: 246-248)
20	Fragebogen zur Evaluation von Lernveranstaltungen (Webler, 2005: 66-68)	überwiegend fünfstufig, *von trifft voll zu* bis *trifft gar nicht zu*, mit einer sechsten Möglichkeit *macht hier keinen Sinn*	73	-	-
21	Prozess- und ergebnisorientierte Lehrveranstaltungsevaluation (PELVE) (Loßnitzer et al., 2007)	fünfstufig, endpunktskaliert, von *stimme nicht zu* bis *stimme zu*, mit einer sechsten Möglichkeit *k.a.*	40	1) Arbeitsaufwand, 2) Zufriedenheit, 3) Kompetenzerwerb, 4) Lehrendenverhalten, 5) Studierendenverhalten, 6) Rahmenbedingungen (Loßnitzer et al., 2007: 331)	-
22	Kurzskala zur Lehrvaluation (Zumbach et al., 2007: 6-7)	fünfstufig, von *trifft voll zu* bis *trifft gar nicht zu*	15	1) Form und Struktur, 2) Merkmale des Dozenten, 3) Umfang und Relevanz, 4) Lernerfolg	- Cronbachs α von .58 bis 73 (Zumbach et al., 2007: 7)
23	Frankfurter-Studierendenfragebogen zur Evaluation von Lehrveranstaltungen (STUD-FEL) (Moosbrugger/Krömker, 2011)	sechsstufig, endpunktskaliert, von *trifft nicht zu* bis *trifft zu*	20	1) Strukturierung und Klarheit, 2) verständliche Erklärungen und Darstellungen, 3) Zeit und Schwierigkeitsmanagement, 4) Verarbeitungstiefe, 5) Motivierung, 6) Kooperation und Klima, 7) Interaktion (Förderung, Leitung, Feedback) (Tillmann et al., 2011: 81)	- Cronbachs α der 12 Items auf Personenebene von .49 bis .76 - Cronbachs α der 12 Items auf Veranstaltungsebene von .42 bis .88

					- Cronbachs α der Gesamtskala auf Personenebene = .91 - Cronbachs α der Gesamtskala auf Veranstaltungsebene = .94 (Tillmann et al., 2011: 84) - Cronbachs α der Gesamtskala = .91 (Peters et al., 2011: 342)
24	Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE und HILVE II) (Rindermann, 2009: 385-391; Rindermann/Amelang, 1994: 61-63)	siebenstufig, von *trifft nicht zu* bis *trifft völlig zu*	32 HILVE; 51 HILVE II	15 Dimensionen, die sich auf 4 Faktoren verteilen; 1) Lehrverhalten, 2) Rahmenbedingungen, 3) studentisches Verhalten, 4) Lehreffektivität (Rindermann, 2009: 90-92)	<u>HILVE:</u> - Cronbachs α für die Faktoren auf Personenebene von .75 bis. 86 - Cronbachs α für die Faktoren auf Veranstaltungsebene von .84 bis. 94 <u>HILVE II:</u> - Cronbachs α für die Faktoren auf Personenebene von .40 bis. 84 - Cronbachs α für die Faktoren auf Veranstaltungsebene von .53 bis. 96 (Rindermann, 2009: 140-141)
25	Studentische Modulevaluation (StudMod) (Lück, 2009: 255-259)	fünfstufig, endpunktskaliert, von *trifft nicht zu* bis *trifft zu*	30 Teilmodulskala; 30 Modulskala	<u>Teilmodulskala:</u> 1) Umgang des Dozierenden mit den Studierenden, 2) Förderung von Verstehen, 3) Strukturierung der Veranstaltung, 4) Einbindung der Studierenden, 5) Förderung der	<u>Teilmodulskala:</u> - Cronbachs α für die Faktoren von .76 bis .82 <u>Modulskala:</u>

				intrinsischen Motivation der Studierenden <u>Modulskala:</u> 1) Modalitäten der Prüfungsleistungen, 2) Rahmenbedingungen, 3) Abstimmung der Teilmodule, 4) Voraussetzungen, 5) Umfang der Prüfungsvorbereitung (Lück, 2009: 47-48)	- Cronbachs α für die Faktoren von .68 bis .80 (Lück, 2009: 91-92).
26	Münsteraner Fragebogen zur Evaluation von Vorlesungen – Revidiert (MFE-VR) (Thielsch/Hirschfeld, 2012: 1-2)	siebenstufig, vollständig beschriftet, von *stimme gar nicht zu* bis *stimme vollkommen zu*	12	1) Dozent und Didaktik, 2) Überforderung, 3) Materialien	- Cronbachs α für die Faktoren von .81 bis .93 (Thielsch/Hirschfeld, 2012: 5)
27	Kurzes Inventar zur Evaluation durch Studierende (KIES) (Fischer, 2014)	variierend, endpunktskaliert (Fischer, 2016, persönliche Mitteilung)	13 in einer Fassung von 08/2016	1) Lehrpersonal, 2) Lehr- und Lernmaterialien, 3) Organisation der Lehrveranstaltung, 4) Inhalte der Lehrveranstaltung, 5) Patientenbezug, 6) Prüfung (Fischer, 2014: 44)	-
28	Das Inventar zur Evaluation von Blended Learning (IEBL) (Peter et al., 2014a: 1-5)	siebenstufig, endpunktskaliert, von *trifft nicht zu* bis *trifft völlig zu*	46	1) Allgemeiner Nutzen, 2) didaktische Qualität, 3) Angemessenheit, 4) Akzeptanz (online), 5) fehlender sozialer Austausch, 6) Usability, 7) Akzeptanz (Präsenz), 8) Dozent (Peter et al., 2014b)	- Cronbachs α für die Faktoren von .69 bis .89 (Peter et al., 2014b)
29	Fragebogen zur allgemeinen Lehrveranstaltungsbewertung (Fondel et al., 2014)	sechsstufig, von *trifft gar nicht zu* bis *trifft voll zu*	verschiedene Arbeitsversionen (Weis et al., 2014)	1) Struktur und Didaktik, 2) Anregungsgehalt, 3) Veranstaltungsklima, 4) Anwendungs- und Praxisbezug (Fondel et al., 2014: 127)	- Reliabilität für die Faktoren von .60 bis .94 (Fondel et al., 2014: 131)
30	Berliner Evaluationsinstrument für Vorlesungen (BLEI-VL) (Landes/Ziegler, 2015: 140)	sechsstufig, von *stimmt gar nicht* bis *stimmt sehr* (Landes/Ziegler, 2015: 138)	13	1) Struktur und begleitende Materialien, 2) Ausgestaltung/Didaktik, 3) Interaktion	- McDonald´s ω für die Faktoren von .89 bis .96 (Landes/Ziegler, 2015: 141)

3 Methodik

In diesem Abschnitt wird das methodische Vorgehen zur Prüfung der aufgeführten Fragebögen (vergleiche Tabelle 1) beschrieben. Auf Ebene der Fragebögen wurde eine Prüfung hinsichtlich relevanter Haupt- und Nebengütekriterien durchgeführt. Zusätzlich wurde auf Ebene der einzelnen Items eine inhaltlich-methodische Prüfung vollzogen. Im Ergebnis erfüllt keiner der aufgeführten Fragebögen zur Evaluation von Lehrveranstaltungen die für den Einsatzweck in der Abteilung Betriebswirtschaft relevanten Haupt- und Nebengütekriterien vollständig. Auch die inhaltlich-methodische Prüfung auf Ebene der Items offenbarte Mängel bei einigen der analysierten Fragebögen, im Schwerpunkt bei der Formulierung. In Konsequenz der Prüfung der aufgeführten Fragebögen wurde ein neuer *Evaluationsfragebogen zur Erfassung studentischer Lehrurteile* (im Folgenden nur noch *EEsL* genannt) kreiert (vergleiche *4 Ergebnis*). Das Vorgehen der Item-Genese des *EEsL* wird nachfolgend dargestellt.

Zunächst wurde nach deutschsprachigen Fragebögen zur Evaluation von Lehrveranstaltungen in der einschlägigen wissenschaftlichen Fachliteratur recherchiert. Anschließend wurden die identifizierten Fragebögen zur Evaluation von Lehrveranstaltungen hinsichtlich ihrer Haupt- und Nebengütekriterien (Bühner, 2011: 58-76) bezüglich eines möglichen Einsatzes in der Evaluation von Lehrveranstaltungen der Abteilung Betriebswirtschaft beurteilt. Als Hauptgütekriterium wird im Folgenden insbesondere das Maß der internen Konsistenz betrachtet. *Objektivität* – unterteilt in Durchführungs-, Auswertungs- und Interpretationsobjektivität – wird nicht geprüft, da alle recherchierten Fragebögen, sieht man einmal von offenen Fragen ab, die Anforderungen der Auswertungs- und Interpretationsobjektivität erfüllen. Durchführungsobjektivität sollte bei jeder standardisierten Lehrevaluation vorliegen. Bereits heute wird die Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft standardisiert durchgeführt. Das Gütekriterium *Validität* wird nachfolgend ebenfalls nicht näher betrachtet, da bis dato ein Dissens hinsichtlich der relevanten Prädiktoren von *Lehrqualität* besteht, der in Kombination mit meist sehr umfangreichen Ansätzen zur Bestimmung von Lehrqualität eine theoriebasierte Entwicklung eines ökonomischen Fragebogens zur Evaluation von Lehrveranstaltungen schwierig gestaltet. Ein Konsens besteht bis dato lediglich hinsichtlich der Mehrdimensionalität von *Lehrqualität*, nicht aber, welche Dimensionen beinhaltet sein sollten (Böttcher/Grewe, 2010: 74; Pohlenz/Oppermann, 2010: 7). Zwei fundierte und überzeugende Ansätze zur Bestimmung relevanter Prädiktoren von *Lehrqualität* – die allerdings zur Konzeption eines ökonomischen Fragebogens zu umfangreich waren – seien aber kurz erwähnt.

Einen umfangreichen Versuch eines theoretischen Ansatzes zur Bestimmung von *Lehrqualität* legte Koch (2004) vor. Anhand von kognitionspsychologischen und instruktionspsychologischen Erkenntnissen sowie konstruktivismusbezogenen Lehrkonzeptionen leitete Koch (2004: 217-225) ein *Ziel-Mittel-Modell* guter und effektiver Lehre ab, das inhaltlich aus 20 Merkmalen besteht (Koch, 2004: 378-381). Ähnlich umfangreich ist Rindermanns (2009: 66) *multifaktorielles Modell der Lehrveranstaltungsqualität*, das theoriebasiert und empirisch hergeleitet wurde und im Rahmen dieser Arbeit zunächst als Grundlage für die Einordnung der Items der aufgeführten Fragebögen diente. Es diente nur als Grundlage, da es für die Konzeption des *EEsL* mit inhaltlich mehr als 20 Bereichen zu umfangreich war (vergleiche Rindermann, 2009: 66). Auch Kochs (2004: 217-225) Modell wurde als zu umfangreich beurteilt und erfüllt damit ein wichtiges Nebengütekriterium (hier: *Ökonomie*) für den Einsatz in der Abteilung Betriebswirtschaft ebenfalls nicht.

Als Nebengütekriterien nach Bühner (2011: 71-76) wurden insbesondere eine mögliche *Normierung*, *Ökonomie* und *Zumutbarkeit* berücksichtigt, wobei die beiden letzteren in engem in-

haltlichen Zusammenhang stehen können. Ein Fragebogen zur Evaluation von Lehrveranstaltungen sollte einerseits ökonomisch sein, um möglichst wenige Ressourcen zu beanspruchen, und andererseits zumutbar, um eine möglichst hohe Motivation zum Ausfüllen bei der Zielgruppe zu gewährleisten. *Zumutbarkeit* wird hier mit dem Umfang eines Fragebogens verknüpft und nicht mit der Operationalisierung sensitiver Fragen, wie sie beispielsweise bei Befragungen zu sexuellen Neigungen oder kriminellen Handlungen auftreten können. Aus der Forschung zu Evaluationen von Lehrveranstaltungen ist bekannt, dass das mehrmalige Ausfüllen eines Fragebogens innerhalb eines kurzen Zeitraums zu Motivationsdefiziten führen kann (Diehl/Kohr, 1977: 74; Gediga et al., 2000: 56). Mindestens einmal im Jahr müssen gemäß Paragraph 5 Absatz 1 Satz 1 des NHG Lehrveranstaltungen evaluiert werden, was dazu führt, dass Studierende innerhalb eines kurzen Zeitraums gebeten werden, die von ihnen besuchten Lehrveranstaltungen zu bewerten und folglich eine gewisse Demotivation die Folge sein kann, wenn beispielsweise ein unnötig langer Fragebogen verwendet wird und immer wieder ausgefüllt werden soll. Reissert (1992) spricht in dem Zusammenhang gar von der *„Kunst"* (Reissert, 1992: 21), nur wenige Items zu verwenden.

Nicht berücksichtigt wurden die Nebengütekriterien *Vergleichbarkeit*, *Nützlichkeit*, *Fairness* und *Unverfälschbarkeit* da die Nebengütekriterien nach Bühner (2011: 71-76) überwiegend der berufsbezogenen Eignungsdiagnostik entstammen und folglich nicht vollständig auf die Auswahl oder Konzeption eines Fragebogens zur Evaluation von Lehrveranstaltungen übertragbar sind. Das Nebengütekriterium *Vergleichbarkeit* (Bühner, 2011: 72) beispielsweise bezieht sich auf Parallelformen von Tests, wie sie aus der Leistungsmessung bekannt sind (vergleiche Lehrl, 2005: 20 mit den *Mehrfachwahl-Wortschatztests A* und *B*) und in der Evaluation von Lehrveranstaltungen keine Anwendung finden (vergleiche Döring und Bortz, 2016: 466-467 zur Erstellung von Parallelformen). *Fairness* hingegen bezieht sich auf das Vermeiden einer möglichen Diskriminierung von Testpersonen innerhalb eines Tests, die beispielsweise aufgrund ethnischer Spezifika einzelner Gruppen (Kubinger/Proyer, 2010: 178) entstehen kann. Somit ist *Fairness* ein weiteres Nebengütekriterium, das im Zusammenhang der berufsbezogenen Eignungsdiagnostik besonders wichtig ist, aber im Rahmen der Evaluation von Lehrveranstaltungen eher vernachlässigbar. *Unverfälschbarkeit* blieb unberücksichtigt, da der Befragungsgegenstand – beispielsweise die *Struktur* einer Lehrveranstaltung oder die *Lehrleistung* eines Lehrenden – in allen recherchierten Fragebögen offensichtlich erkennbar war. Möchte ein Studierender einen Lehrenden beispielsweise aufgrund persönlicher Rachemotive bei schlechtem Abschneiden in einer Klausur negativ evaluieren, wird ihm das aufgrund des eindeutigen Befragungsgegenstands in der Regel auch gelingen. Zwar ließen sich auch aufwändig konzipierte Testverfahren zur Evaluation von Lehrveranstaltungen kreieren, die weniger verfälschbar wären als Fragebögen, gleichzeitig ginge solch ein Vorgehen aber zulasten einer ökonomischen Konzeption und einer ökonomischen Durchführung und Auswertung, die hier höher gewichtet waren als die *Unverfälschbarkeit*. *Unverfälschbarkeit* ist folglich ein sinnvolles Gütekriterium der berufsbezogenen Eignungsdiagnostik, beispielsweise um Selbstdarsteller zu enttarnen, die sich zu positiv darstellen, aber kein sinnvolles Gütekriterium eines guten Fragebogens zur Evaluation von Lehrveranstaltungen, insbesondere da der Befragungsgegenstand offensichtlich sein sollte um verlässliche Bewertungen zu erzielen. Das Nebengütekriterium *Nützlichkeit* war zunächst nicht relevant, da es sich einerseits auf das praktische Bedürfnis einer Messung bezieht – was strenggenommen qua Gesetz vorgegeben ist – und andererseits auf die Neukonzeption von Fragebögen. Neu konzipierte Fragebögen sollten gemäß dem Nebengütekriterium *Nützlichkeit* einen Vorteil gegenüber bereits bestehenden Testverfahren mit gleichem Ziel der Messung aufweisen (Bühner, 2011: 73-74). Wie im Folgenden noch dargestellt wird, wurde das Nebengütekriterium *Nützlichkeit* erst im Verlauf des analytischen Prozesses relevant.

Tabelle 2 veranschaulicht, welche Gütekriterien berücksichtigt wurden. Wichtig sei der Hinweis, dass die Unterteilung in Haupt- und Nebengütekriterien auf Bühner (2011: 58-74) zurückzuführen ist. Im Rahmen dieser Arbeit stehen Haupt- und Nebengütekriterien gleichberechtigt nebeneinander.

Tabelle 2: Relevante und irrelevante Gütekriterien (eigene Darstellung).

Gütekriterium	Relevanz	Begründung
		Hauptgütekriterien
Reliabilität	ja	nur zuverlässige Fragebögen sollen zur Auswahl herangezogen werden
Objektivität	nein	Fragebögen mit geschlossenem Antwortformat erfüllen Objektivität
Validität	nein	Konzepte zur Lehrqualität sind sehr umfangreich und konterkarierten bei Berücksichtigung das Kriterium der Ökonomie
		Nebengütekriterien
Normierung	ja	Normwerte ermöglichen eine Vergleichbarkeit, beispielsweise mit anderen Fachbereichen
Ökonomie	ja	aufgrund der hohen empirischen Zusammenhänge inhaltlich unabhängiger Bereiche, soll ein neuer Fragebogen ökonomischer messen
Zumutbarkeit	ja	es besteht ein enger inhaltlicher Zusammenhang zur Ökonomie
Nützlichkeit	ja	bezieht sich hier auf den Mehrwert eines neu konzipierten Fragebogens gegenüber bereits bestehenden Fragebögen mit demselben Ziel der Messung
Vergleichbarkeit	nein	erfordert Parallelformen von Tests, die im Rahmen der Evaluation von Lehrveranstaltungen nicht gängig sind
Fairness	nein	bezieht sich auf eine mögliche Diskriminierung von einzelnen Gruppen innerhalb eines Fragebogens
Unverfälschbarkeit	nein	sinnvoll im Rahmen der berufsbezogenen Eignungsdiagnostik um Selbstdarsteller zu enttarnen, weniger sinnvoll im Rahmen der Evaluation von Lehrveranstaltungen, wo der Befragungsgegenstand offensichtlich ist

Die Fragebögen (vergleiche Tabelle 1) zur Evaluation von Lehrveranstaltungen wurden anhand der relevanten Gütekriterien beurteilt. Aufgrund der methodischen Mängel, die bei der Itemformulierung des derzeit in der Abteilung Betriebswirtschaft verwendeten Fragebogens auffielen, wurden zusätzlich alle Items der recherchierten Fragebögen hinsichtlich methodischer Mängel überprüft. Die Überprüfung aller Items führte – in Kombination mit den relevanten Gütekriterien – zu dem Ergebnis, dass keiner der recherchierten Fragebögen für eine Verwendung in der Abteilung Betriebswirtschaft unmittelbar infrage kam. Aus diesem Schritt resultierte die Berücksichtigung des Gütekriteriums *Nützlichkeit*.

Die Beurteilung der aufgeführten Fragebögen als unzureichend mag aufgrund der Vielzahl (vergleiche Tabelle 1) vermessen klingen, soll aber in keiner Weise die jeweiligen Autoren kritisieren, denn den geprüften Fragebögen lagen bei der Konstruktion vermutlich andere Ziele zugrunde als die mit dieser Ausarbeitung verfolgten. Das kann am Beispiel des von Rindermann und Amelang (1994: 61-63) entwickelten *Heidelberger Inventar zur Lehrveranstaltungsevalu-*

ation (kurz: *HILVE*) veranschaulicht werden. Der *HILVE* wurde umfassend empirisch überprüft, in einer zweiten verbesserten Version vorgelegt und entspricht den Anforderungen, beispielsweise den psychometrischen Gütekriterien, die an einen Fragebogen zu stellen sind (Rindermann/Amelang, 1994: 13-24; Rindermann, 2009: 133-134). Problematisch für den Einsatz in der standardisierten regelmäßigen Evaluation von Lehrveranstaltungen der Abteilung Betriebswirtschaft ist jedoch die Länge des *HILVE* (Rindermann, 2009: 385-391). Denn eine in die Überprüfung – des aktuell in der Abteilung Betriebswirtschaft eingesetzten Fragebogens – einbezogene studentische Expertengruppe befand bereits den aktuell verwendeten Fragebogen als zu lang, obwohl er wesentlich kürzer (28 Items versus 51 Items) als beispielsweise der *HILVE-II*-Fragebogen (Rindermann, 2009: 388-391) ist.

Da die Idee der unveränderten Übernahme eines bestehenden Fragebogens aufgrund der dargelegten Gründe im Verlaufe des Analyseprozesses verworfen werden musste, wurden die Items der aufgeführten Fragebögen inhaltlich gruppiert und einzeln hinsichtlich ihrer Eignung überprüft. Das Ziel der Gruppierung war es, inhaltliche Schwerpunkte zu identifizieren, damit anschließend methodisch überzeugende Items zur Konzeption des *EEsL* verwendet werden können. Ein Vorgehen, dass Westermann, Spies, Heise und Wollburg-Claar (1998: 140) sogar empfehlen, denn ihrer Ansicht nach existiere bereits ein ausreichend großer Pool an Items, der zur vollständigen Erfassung von *Lehrqualität* verwendet werden könne (vergleiche auch Reissert, 1992: 21). Und obwohl es aus testtheoretischer Sicht vorteilhafter wäre einen geeigneten bestehenden Fragebogen unverändert einzusetzen, anstelle von einer Kombination an Items verschiedener Fragebögen, wurde aufgrund der dargelegten Gründe dieses Vorgehen gewählt.

Die Itemanalyse erfolgte schrittweise. Zunächst wurden die Items der in Tabelle 1 aufgeführten Fragebögen händisch in eine Excel-Tabelle übertragen und inhaltlich nach Rindermann (2009: 66) gruppiert. Insgesamt wurden in diesem Analyseschritt rund 1.200 Items berücksichtigt. Anschließend wurden alle Items in einer Gesprächsrunde geprüft. Die Gesprächsrunde bestand aus zwei Lehrenden der Hochschule Hannover, die Forschungsmethoden im Masterstudiengang Unternehmensentwicklung lehren und mit den Anforderungen an eine Fragebogenkonstruktion vertraut sind. Aussortiert wurden beispielsweise Items wie „Der Dozent ist gut vorbereitet“ oder Fragen wie „Ist der Dozent fachlich kompetent?“, da sie für Studierende nicht eindeutig beurteilbar sind (Diehl, 2003: 40; Koch, 2004: 27; 227; Lück, 2005: 184). Ebenso können einige Aspekte von *Lehrqualität* durch Studierende nur schwer beantwortet werden, beispielsweise die Beurteilung der wissenschaftlichen Qualität von Lehrinhalten (Rindermann/Amelang, 1994: 11; Westermann et al., 1998: 135). Ähnliches stellen auch Metje und Kelle (2010: 102-104) fest. Mittels qualitativer Interviews – unter anderem zum *Fragebogen zur Lehrveranstaltungsevaluation von Vorlesungen* (kurz: *FEVOR*) (Staufenbiel, 2000: 178-179) – identifizieren Metje und Kelle (2010: 102-104) Items wie „Die Vorlesung/das Seminar gibt einen guten Überblick über das Themengebiet“ oder „Das Seminar/Die Vorlesung ist vermutlich für die spätere Berufspraxis sehr nützlich“ als schwierig zu beurteilen für Studierende. Und obwohl diese methodischen Hinweise bereits länger bekannt sind (vergleiche Brehl/Schneider/Utsch, 2004: 317, el Hage, 1996: 101-105 oder Reischmann, 1995: 271-272 für eine Übersicht), wurden noch viele Items mit aus methodischer Perspektive zweifelhaften Formulierungen gefunden und aussortiert. Wohl auch, weil im Rahmen der Evaluation von Lehrveranstaltungen nicht immer ein Konsens bezüglich eines methodisch einwandfreien Vorgehens zur Formulierung von Items vorzuliegen scheint. Schafferer (2010: 67) kritisiert beispielsweise an Rindermanns (2009: 66) *multifaktoriellem Modell der Lehrveranstaltungsqualität* den fehlenden Aspekt der wissenschaftlichen Aktualität von Lehrinhalten, obwohl ihn, auch im Verständnis des Autors, Studierende zum Zeitpunkt der Evaluation nur schwer einschätzen können. Andere Items wurden aussortiert, weil sie inhaltlich nicht die *Lehrqualität*

erfassen, beispielsweise „Die Lehrperson konnte mein Interesse an den Veranstaltungsinhalten wecken“. Bei Items wie diesem ist nicht klar, ob bei einer eher negativen Beantwortung die Lehrleistung eines Lehrenden ungenügend oder das Desinteresse des Evaluierenden schlicht zu groß war. Anekdotisch seien zudem einige Items genannt, deren Formulierungen Zweifel an einer sinnvollen Erhebung weckten: „In der Lehrveranstaltung ging es im Großen und Ganzen sehr lustig zu“, „Der Dozent suchte auch persönlichen Kontakt mit den Teilnehmern“, „Das hier Gelernte kann man später sicher irgendwie anwenden“ oder „Diese Lehrveranstaltung ist langweilig und einschläfernd“. Wird bei den ersten drei Beispielen nicht genau klar, was die Items inhaltlich erfassen sollen, erscheint das letzte Beispiel insofern ungeeignet, als es negativ formuliert ist und in der Rückmeldung Demotivation begünstigen kann.

Darüber hinaus wurden Items aussortiert, die das *Zeitmanagement* eines Lehrenden oder das *Schwierigkeitsniveau* einer Lehrveranstaltung operationalisieren. Sowohl Fragen zum *Zeitmanagement* als auch Fragen zum *Schwierigkeitsniveau* operationalisieren inhaltlich häufig die subjektiv empfundene Angemessenheit, deren Beurteilung nach Ansicht des Autors aber unter anderem mit *Intelligenz* konfundiert ist. Insbesondere aufgrund stark gestiegener Studienanfängerquoten (BMBF, 2016: 44) ist von Fragen, die mit *Intelligenz* konfundiert sein könnten, abzusehen, da davon auszugehen ist, dass, wenn immer mehr Personen eines Geburtsjahrgangs ein Studium beginnen, auch vermehrt weniger talentierte Personen ein Studium aufnehmen und nicht mehr nur die talentiertesten eines Geburtsjahrgangs. Eine Studienanfängerquote von rund 58 Prozent in den Jahren 2014 und 2015 (33 Prozent im Jahr 2000) (BMBF, 2016: 44) verdeutlicht diese Bedenken, veranschaulichen die 58 Prozent doch, dass selbst bei einer normalverteilten Variable und der Voraussetzung – nur die talentiertesten Personen eines Geburtsjahrgangs begännen ein Studium – auch Personen ein Studium starteten, die unterhalb des Durchschnitts lägen. *Intelligenz* ist eine Variable, die bei ausreichend großer Stichprobe annähernd normalverteilt ist (Stern/Grabner, 2014: 178), weshalb die Konfundierung von *Intelligenz* mit subjektiv beurteilter Angemessenheit des *Zeitmanagements* oder des *Schwierigkeitsniveaus* bedacht werden sollte. Auch vor dem Hintergrund einer zu milden Beurteilung von Studierenden durch Lehrende sei dieser Aspekt erwähnt, könnte doch die häufige Rückmeldung zu schwer empfundener Lehrinhalte auch die Konzeption einer Lehrveranstaltung beeinflussen, dahingehend, sie leichter zu gestalten. Ein Aspekt, der im Zusammenhang einer zyklischen Verbesserung von Abschlussnoten innerhalb der Fachdisziplin Betriebswirtschaftslehre zwar nicht zwingend für eine vollständige retrospektive Erklärung infrage kommt (Gaens, 2015: 14-20; 26), eine prospektive weitere Verbesserung aber – nach Ansicht des Autors – begünstigen kann. Dass eine leistungsunabhängige Verbesserung von Abschlussnoten – insbesondere einhergehend mit einer verringerten Varianz hinsichtlich der Differenzierbarkeit – beispielsweise im Rahmen von Personalauswahlprozessen problematisch ist, scheint unzweifelhaft. Bislang sind die genauen Ursachen für eine Verbesserung von Abschlussnoten im Zeitverlauf nicht eindeutig geklärt, vergleiche aber Gaens (2015: 25-29) für mögliche Erklärungsansätze.

Aus den dargelegten Gründen wurde bei der Konzeption des *EEsL* auf Fragen zum *Schwierigkeitsniveau* und zum *Zeitmanagement* verzichtet. Berücksichtigt wurde *Intelligenz* zudem bei der Auswahl und Formulierung von Items. Insbesondere empirische Befunde zur *sprachlichen Intelligenz* sollten konzeptionell bei der Formulierung von Items berücksichtigt werden. Die Ergebnisse mehrerer Studien (Heber, 2013; Linssen/Mayer, 2016: 43-44; Litzcke/Linssen/Maffenbeier/Schilling, 2012: 112; Schön, 2011: 106), bei denen ein Intelligenztest (Lehrl, 2005: 44-45) zum Einsatz kam, mittels dessen die *sprachliche Intelligenz* operationalisiert wird und anhand der auf einen allgemeinen Intelligenzfaktor geschlossen werden kann, stimmen nachdenklich. Unabhängig von der verwendeten Stichprobe – Berufsschüler oder Studierende

– wurden unterdurchschnittliche Werte erzielt, obwohl beispielsweise Studierende als Angehörige der Bildungselite, verglichen mit der Allgemeinbevölkerung, eher überdurchschnittliche Resultate erzielen sollten. Rindermann, Baumeister und Gröper (2014: 206) diagnostizierten beispielsweise an einer Gruppe Ingenieure (n=30, 77 Prozent männlich) – bei der Teile eines Intelligenztests verwendet wurden, die mathematische und figurale Fähigkeiten messen – einen durchschnittlichen Intelligenzquotienten in Höhe von 116, somit mehr als eine Standardabweichung besser als der eigentliche Durchschnitt (100) in der Allgemeinbevölkerung. Nun lässt sich zwar einwenden, Studierende eines Ingenieursstudiengangs wären aufgrund einer mathematischen Affinität bevorteilt, gleiches sollte aber auch für Studierende gelten, deren Lehrinhalte zu einem beträchtlichen Teil Textverständnis erfordern. In gewisser Hinsicht wären bei Studierenden der sozialen Arbeit (Linssen/Mayer, 2016: 43-44; Litzcke/Linssen/Maffenbeier/Schilling, 2012: 112) höhere Werte hinsichtlich der sprachlichen Intelligenz zu erwarten gewesen, da die spätere berufliche Tätigkeit oftmals Aufgaben beinhaltet, die sprachliche Intelligenz verlangt (Linssen/Mayer, 2016: 43). Ähnliches gilt auch für Teile der Studierenden der Betriebswirtschaftslehre (Heber, 2013; Litzcke/Linssen/Maffenbeier/Schilling, 2012: 112). Zwar können Studierende der Betriebswirtschaftslehre an der Hochschule Hannover Schwerpunkte wie Corporate Finance und Controlling wählen, die vermutlich aufgrund eines hohen mathematischen Bezugs weniger sprachliche Intelligenz erfordern als andere Schwerpunkte, nichtsdestotrotz bilden die genannten Schwerpunkte nur einen Teil des Studiums ab. Am Ende des Studiums müssen die Studierenden beispielsweise eine Bachelorarbeit anfertigen, die unzweifelhaft sprachliche Intelligenz erfordert.

Die Resultate im Bereich der *sprachlichen Intelligenz* von Studierenden der sozialen Arbeit (Linssen/Mayer, 2016: 43-44; Litzcke/Linssen/Maffenbeier/Schilling, 2012: 112), der Betriebswirtschaftslehre (Heber, 2013; Litzcke/Linssen/Maffenbeier/Schilling, 2012: 112) sowie der Lehrberufe Bankkaufmann, Kaufmann für Versicherungen und Finanzen sowie Sozialversicherungsfachangestellter (Schön, 2011: 106) waren unerwartet. Und wenngleich der Beitrag von Linssen und Mayer (2016) kritische Repliken hervorrief (Schone, 2016: 122; van Randenborgh, 2016: 123), sind ihre Befunde mit dem *Mehrfachwahl-Wortschatz-Intelligenztest B* (kurz: *MWT-B*) (Lehrl, 2005: 44-45) konsistent. Insofern erstaunen die Zweifel an Linssens und Mayers (2016: 43-44) Befunden, wonach das Sprachniveau von Studierenden der Sozialen Arbeit unterdurchschnittlich ist. Zumal ein Teil der Kritik auch hinfällig geworden wäre, hätte man die aktuelle empirische Befundlage recherchiert und in die Beurteilung einbezogen. Insbesondere hinsichtlich der Kritik an möglicherweise veralteten Normwerten erscheint van Randenborghs (2016: 123) Replik voreilig verfasst, denn sie basiert lediglich auf Vermutungen, obwohl die Empirie fundierte Befunde bereithält. Ihre Zweifel (van Randenborgh, 2016: 123) an möglicherweise veralteten Normwerten und die Kritik an den lediglich moderaten Zusammenhängen mit ausführlicheren Intelligenztests (vergleiche auch Lehrl, 2005: 24) sind nachvollziehbar, interessanterweise führt Satzgers, Fessmanns und Engels (2002: 161-166) Überprüfung hinsichtlich der Normenäquivalenz verschiedener Intelligenztests aber zu dem Befund, dass die *sprachliche Intelligenz* mittels *MWT-B* (Lehrl, 2005:44-45) eher überschätzt wird. Im Detail erzielten gesunde Testpersonen mit dem *MWT-B* (Lehrl, 2005: 44-45) einen um rund 17 Punkte höheren sprachlichen Intelligenzquotienten als Testpersonen, die den sprachlichen Teil eines ausführlicheren Intelligenztests bearbeiteten (Satzger/Fessmann/Engel, 2002: 168).

Auf weitere empirische Befunde sei an dieser Stelle verzichtet, denn für diese Arbeit sind die Zusammenhänge zwischen einzelnen Intelligenztests nicht zwingend erforderlich. Dem interessierten Leser sei aber eine Übersichtsarbeit von Antretter, Dunkel und Haring (2013) emp-

fohlen, die systematisch deutschsprachige Leistungstests aus der Erwachsenenpsychiatrie hinsichtlich ihres theoretischen Hintergrunds, ihrer Aktualität der Normwerte und ihrer psychometrischen Gütekriterien geprüft haben.

Wichtig für diese Arbeit erscheint es, einen möglicherweise im Zeitverlauf gesunkenen verbalen Intelligenzquotienten bei der Item-Genese zu berücksichtigen. Insbesondere die Befunde von Litzcke, Linssen, Maffenbeier und Schilling (2012: 112) sowie Heber (2013) in der Abteilung Betriebswirtschaft veranlassten zur Berücksichtigung und zu besonders verständlich formulierten Items. Hinsichtlich der kritischen Repliken (Schone, 2016: 122; van Randenborgh, 2016: 123) auf Linssens und Mayers (2016) Beitrag ist jedoch Wieland (2016: 124) zu unterstützen, der die Frage nach der Generalisierbarkeit der aktuellen Befunde empirisch klären möchte. Insbesondere die Replik von Schone (2016) lässt befürchten, dass der Diskurs nicht nur nach wissenschaftlichen Kriterien erfolgen könnte (vergleiche auch Linssen, 2016: 156). Das sollte jedoch nicht davon abhalten, die sinkenden Durchschnittswerte in der Variable *Intelligenz* bei Studierenden ernst zu nehmen und inhaltlich zu reagieren.

Auffällig bei der Analyse der Items war, dass ein Sättigungseffekt eintrat und Items verschiedener Fragebögen inhaltlich ähnliche Formulierungen aufwiesen. Ähnlich formulierte Items fanden sich beispielsweise zur *Struktur* einer Lehrveranstaltung oder zur *Lehrleistung* eines Lehrenden. Zumbach et al. (2007: 8) mutmaßen, dass Fragebögen zur Evaluation von Lehrveranstaltungen aufgrund eines ähnlichen Vorgehens bei der Genese der Items homogen sind. Ein Großteil der konzipierten Fragebögen fußt nicht etwa auf einem diagnostischen Konstruktionsprinzip, sondern wurde ad-hoc erstellt (Gollwitzer/Kranz/Vogel, 2006: 90).

In einem zweiten Schritt wurden diejenigen Items gruppiert und analysiert, die sich innerhalb des ersten Schritts als verwendbar erwiesen hatten. Die Anzahl der Items war bei diesem Schritt bereits kleiner als noch im ersten Schritt. Insgesamt wurden im zweiten Schritt 99 Items berücksichtigt. Gruppiert wurden die Items nach Schmidt und Loßnitzer (2010: 58), deren inhaltlicher Fokus auf den Merkmalen Input, Prozess, Output und Qualitätsentwicklung liegt und die wiederum noch Subdimensionen beinhalten. Der Wechsel des Kategoriensystems von Schritt 1 zu Schritt 2 geschah aufgrund der verbesserten Übersicht durch Schmidts und Loßnitzers (2010: 58) Modell. Verteilten sich die rund 1.200 Items im ersten Schritt noch auf 18 Kategorien, reduzierte sich die Anzahl im zweiten Schritt auf 10 Kategorien bei 99 Items. Aussortiert wurden im Rahmen des zweiten Schritts unter anderen solche Items, die erst bei näherer Betrachtung ungeeignet erschienen, beispielsweise Items, die Erwartungen operationalisieren. „Es finden ausreichend Diskussionen statt“ ist ein Beispiel für ein Item, das Erwartungen operationalisiert. Werden Erwartungen operationalisiert, kann eine Konfundierung von Erwartungshaltung und Evaluation die Folge sein (Koch, 2004: 27), was folglich die Validität einer Messung infrage stellt.

Da Fragebögen zur Evaluation von Lehrveranstaltungen oft überwiegend Items beinhalten, die lehrendenzentriert formuliert sind, wurden zudem einige dieser Items aussortiert um den inhaltlichen Fokus nicht zu sehr auf einen Lehrenden zu konzentrieren. Gegen die Verwendung zu vieler lehrendenzentrierter Items spricht, dass es ein falsches Signal senden würde, nämlich, die Selbstverantwortung von Studierenden sei zum Gelingen einer Lehrveranstaltung mit hoher Qualität nicht notwendig (Reissert, 1992: 18; Webler, 2005: 65). Konzeptionell im Fokus standen Items, die inhaltlich beobachtbares *Verhalten von Lehrenden*, die *Interaktion* zwischen Lehrendem und Studierenden, die *Struktur* der Lehrveranstaltung, den *Beitrag* der *Studierenden* und die *Transformation* hin zu eigenständigem Denken operationalisieren. Der konzeptionelle Fokus ergab sich einerseits aus inhaltlichen Überlegungen, andererseits aus den inhaltlichen Schwerpunkten, die innerhalb der aufgeführten Fragebögen gelegt wurden und die durch die

Gruppierung nach Rindermann (2009: 66) sowie Schmidt und Loßnitzer (2010: 58) zutage traten. Neben methodischen Hinweisen der Lehrevaluationsforschung zur Konzeption eines Fragebogens wurden aber auch klassische Güteprinzipien zur Formulierung von Items berücksichtigt. Aussortiert wurden Items mit unverständlichen und komplizierten Formulierungen, Doppeldeutigkeiten, impliziten Erwartungen, nicht beobachtbaren Inhalten, doppelten Verneinungen, negativer Formulierung, Mehrfachaussagen, umständlichen Längen sowie telegrafischen Abkürzungen (Bühner, 2011: 134-138; Saris/Gallhofer, 2007: 87-88; Jonkisz/Moosbrugger/Brandt, 2012: 64-66).

Aufgrund der Vielzahl zu erfassender Fragebögen im Rahmen der Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft, kommt zur Beantwortung der Mehrheit der Items nur ein gebundenes Antwortformat infrage. Verglichen mit einem freien Antwortformat erfüllen gebundene Antwortformate das Gütekriterium der *Ökonomie* besser (Rost, 2004: 61). Gebundene Antwortformate können aber insofern problematisch sein, als sie bestimmte Antwortstile begünstigen können (Bühner, 2011: 111), beispielsweise wenn Testpersonen eher zu den Rändern der Antwortkategorien tendieren. Eine vollständige Lösung für die Problematik von Antwortstilen ist bislang lediglich die Verwendung dichotomer Antwortkategorien (Bühner, 2011: 111; 117), die jedoch für den *EEsL* nicht infrage kam. Dichotome Antwortkategorien kamen deshalb nicht infrage, weil zwar die Gefahr von Antwortstilen vermieden wird, gleichzeitig aber ein Informationsverlust einhergeht, der den Vorteil an dieser Stelle aufwiegt. Zwar ließen sich auch mehr Items generieren um das Problem des einhergehenden Informationsverlusts wettzumachen (Bühner, 2011: 117), jedoch würde solch ein Vorgehen das Gütekriterium *Ökonomie* verletzen. Eine Analyse der Antwortkategorien der aufgeführten Fragebögen in Tabelle 1 ergab, dass die Mehrzahl ein fünfstufiges Antwortformat aufweist. Bei der Mehrzahl der Fragebögen begründeten die Autoren ihr Antwortformat jedoch nicht. Unklar ist, ob den fehlenden Begründungen der Autoren methodische und inhaltliche Überlegungen vorausgingen, oder ob die Antwortkategorien eher zufällig gewählt wurden. Um das eigene Antwortformat fundierter zu begründen, wurden methodische sowie empirische Erkenntnisse berücksichtigt und unter inhaltlichen Überlegungen für die eigene Arbeit interpretiert.

Bühner (2011: 116) legt sich nicht auf ein gerades oder ungerades System von Antwortkategorien fest und empfiehlt vielmehr, für die Maximierung der Reliabilität einer Skala auf ein fünf- oder siebenstufiges Antwortformat zurückzugreifen. Besteht jedoch die Sorge, die Testpersonen könnten Schwierigkeiten bei der Interpretation einer mittleren Antwortkategorie haben, empfiehlt Bühner (2011: 116) ein sechsstufiges Antwortformat. Eine ähnliche Empfehlung gibt auch Rost (2004: 67), der ebenfalls auf Probleme mit einer mittleren Antwortkategorie verweist. Mitunter unterscheiden sich die Testpersonen darin, wie sie eine mittlere Antwortkategorie auffassen. Einige Testpersonen könnten Items für unpassend halten und deshalb eine mittlere Antwortkategorie wählen, andere Testpersonen könnten ihre Antwort verweigern wollen und aus diesem Grund eine mittlere Antwortkategorie wählen (Rost, 2004: 67). Döring und Bortz (2016) fassen die mögliche Problematik einer mittleren Antwortkategorie unter dem Begriff „*Ambivalenz-Indifferenz-Problem*“ (Döring/Bortz, 2016: 249) zusammen. Eine Testperson kann ihrer (Döring/Bortz, 2016: 249) Ansicht nach die mittlere Antwortkategorie entweder gewählt haben weil sie meinungslos ist oder ihre Antwort tatsächlich der mittleren Antwortkategorie entspricht. Wird erwartet, dass Testpersonen vermehrt eine mittlere Antwortkategorie wählen weil sie a) meinungslos sind, b) Schwierigkeiten bei der Interpretation haben, c) ihre eigentliche Antwort verweigern wollen oder d) Items für unpassend halten, empfehlen Döring und Bortz (2016: 249) eine zusätzliche, neutrale Antwortkategorie, beispielweise „keine Angabe“. Zwar böte sich auch eine Lösung mithilfe eines Systems gerader Antwortkategorien an,

allerdings besteht bei einem geraden Antwortformat die Gefahr, dass Testpersonen sich entscheiden müssen. Insbesondere für Testpersonen, deren Einschätzung einer mittleren Antwortkategorie entspricht, sind gerade Antwortkategorien methodisch heikel. Döring und Bortz (2016: 249) empfehlen deshalb ein gerades System von Antwortkategorien dann zu wählen, wenn die Gefahr einer zu intensiven Nutzung der mittleren Antwortkategorie durch die Testpersonen besteht. Da die methodische Fachliteratur keine eindeutigen Hinweise hinsichtlich einer ungeraden oder geraden Anzahl an Antwortkategorien bereithält, wurden zusätzlich inhaltliche Überlegungen einbezogen.

Unter inhaltlicher Perspektive erscheint es a) abwegig, dass Studierende hinsichtlich der Evaluation einer besuchten Lehrveranstaltung meinungslos sind. Beispielsweise können Unzufriedenheit und Zufriedenheit im Rahmen der Evaluation von Lehrveranstaltungen mitgeteilt werden, von denen, je nach Zeitpunkt der Durchführung und Rückmeldung der Ergebnisse, entweder die Evaluierenden selbst profitieren können, oder aber Folgegenerationen von Studierenden. Insbesondere wenn Lehrende die Ergebnisse der Evaluation von Lehrveranstaltungen aktiv rückmelden, mit den Studierenden besprechen und Kritik konstruktiv auffassen, scheint ein hohes Interesse der Studierenden möglich. Hat sich jedoch eine Kultur des Schweigens zu den Ergebnissen der Evaluation von Lehrveranstaltungen etabliert und leben Lehrende das Interesse an den Ergebnissen nicht selbst vor – tun die Evaluation der Lehrveranstaltung schlimmstenfalls öffentlich als lästige Pflicht ab – kann das zu Motivationsdefiziten und Meinungslosigkeit bei Studierenden führen (vergleiche auch Reissert, 1992: 21).

Dass Studierende b) Schwierigkeiten hinsichtlich der Beurteilung von Items haben und in ihrer kognitiven Überforderung hilflos die mittlere Antwortkategorie wählen, erscheint unwahrscheinlich, da, wie oben dargestellt, unter anderem auf sprachlich abstrakt formulierte Items verzichtet wird. Fragebögen zur Messung von *Persönlichkeit* beispielsweise beinhalten mitunter abstrakt formulierte Items, die zudem noch Bereiche der *Persönlichkeit* im Selbsturteil abfragen, mit denen sich Testpersonen eher selten beschäftigen. Schwierig zu beurteilen können Items für Studierende aber dann sein, wenn sie zur Evaluation einer Lehrveranstaltung gebeten werden, die sie bislang nicht oder nur unregelmäßig besucht haben. Ein Problem, dass zwar auftreten kann, aber als zu gering betrachtet wird um konzeptionell bei der Auswahl gerader oder ungerader Antwortkategorien berücksichtigt zu werden.

Gleiches gilt für das mögliche Problem c) der Antwortverweigerung. Es erscheint zwar möglich, dass Antwortverweigerer auftreten können, im Rahmen der Evaluation einer Lehrveranstaltung ist das aber vermutlich ein eher geringes Problem. Einerseits wurde bereits dargelegt, weshalb Studierende an der Evaluation von Lehrveranstaltungen interessiert sein dürften, andererseits drohen aufgrund einer anonymen Erhebungssituation in der Abteilung Betriebswirtschaft keine Repressalien bei kritischer Evaluation, die eine Antwortverweigerung und das Ausweichen auf die mittlere Antwortkategorie nach sich ziehen könnten.

Wenig relevant erscheint darüber hinaus die Gefahr einer Tendenz zur Mitte aufgrund von d) unpassend empfundenen Items. Da eine studentische Expertengruppe – bestehend aus Studierenden, die mit der Fragebogenkonstruktion vertraut waren, und Vertretern des Fachschaftsrats Wirtschaft – in die konzeptionelle Gestaltung des *EEsL* eingebunden waren, ist von einer ausreichend hohen sozialen Validität der Items auszugehen.

Die inhaltlichen Überlegungen (a bis d) rechtfertigen ein ungerades und gerades Antwortformat gleichermaßen. Auch unter methodischen Gesichtspunkten erschienen beide Antwortformate gerechtfertigt. Empfohlen wird deshalb ein fünf-, ein sechs- oder ein siebenstufiges Antwortformat. Von mehr als sieben Antwortkategorien wird abgeraten, da ein Mehr an Antwortkategorien nicht automatisch auch zu mehr Informationen führt (Jonkisz/Moosbrugger/Brandt, 2012: 51). Von weniger als fünf Antwortkategorien wird abgeraten, weil damit eine

verminderte Varianz einherginge, die zwar mit einem Anstieg der Anzahl an Items kompensiert werden könnte, gleichzeitig aber dem Gütekriterium *Ökonomie* widerspräche. Werden ungerade Antwortkategorien verwendet, kann sich zudem die Aufnahme einer zusätzlichen Antwortkategorie – beispielsweise „keine Angabe" – empfehlen (Döring/Bortz, 2016: 249). Je nach Erkenntnisinteresse und Antwortformat kann die Antwortkategorie „keine Angabe" aber auch weggelassen werden.

Empirische Befunde hinsichtlich der geeigneten Bezeichnungen von Antwortkategorien sind allenfalls spärlich vorhanden. Rohrmann (1978) untersuchte vier fünfstufige Urteilsdimensionen dahingehend, inwieweit die Bezeichnungen der Antwortkategorien mit den hinterlegten Zahlencodierungen übereinstimmten. Da Rohrmann (1978: 242-243) aber selbst die Generalisierbarkeit seiner Ergebnisse aufgrund zu geringer Stichprobenumfänge und zu geringer externer Validität für „unzulänglich" (Rohrmann, 1978: 242) erachtet, sei sein Versuch an dieser Stelle nur der Vollständigkeit halber erwähnt. Da sowohl die methodische Fachliteratur als auch die empirische Forschung keine eindeutigen Hinweise bezüglich der Bezeichnungen der Antwortkategorien bereithalten, wurde die Skalierung unter inhaltlicher Überlegung gewählt. Unter *4 Ergebnisse* werden zwei Vorschläge des *EEsL* mit fünfstufigen Antwortkategorien und der Zusatzkategorie „keine Angabe" unterbreitet. Ein Vorschlag (Tabelle 3) weist Antwortkategorien von *stimme nicht zu* bis *stimme zu* auf, ein weiterer Vorschlag (Tabelle 4) umfasst Antwortkategorien von *sehr gut* bis *sehr schlecht*.

Soll zudem die Wahrscheinlichkeit erhöht werden, dass die Antwortkategorien durch Testpersonen als äquidistant erlebt werden, empfiehlt sich die nummerische Unterstützung verbal beschrifteter Antwortkategorien (Mummendey/Grau, 2014: 82). Guilford (1954: 264) verweist im Zusammenhang einer nummerischen Unterstützung von verbalen Antwortkategorien auf die Problematik bipolarer Skalen, die eine Null beinhalten und auch negative Zahlen verwenden. Unter anderem besteht die Gefahr, dass beispielsweise fünfstufige Antwortkategorien von -2 bis 2 als weniger äquidistant erlebt werden als 0 bis 5. Mummendey und Grau (2014: 82-83) erwähnen zudem die Problematik, dass Testpersonen bei bipolaren Antwortkategorien überwiegend positive Antwortkategorien verwenden um sich selbst keine negativen Eigenschaften zuzuschreiben, was folglich die Varianz in den Antworten einschränkt. Studierende evaluieren jedoch Lehrende, und nicht sich selbst, weshalb eine nummerische Unterstützung verbal beschrifteter Antwortkategorien als sinnvoll erachtet wird, insbesondere mithilfe eines unipolaren Zahlenformats. Auch unter Berücksichtigung motivationstheoretischer Befunde ist die Rückmeldung an Lehrende auf Basis unipolarer Antwortkategorien – unabhängig vom jeweiligen Evaluationsergebnis – konstruktiver als bei bipolaren Antwortkategorien, weil weniger demotivierende negative Ergebnisse resultieren. Diskutabel ist jedoch, ob sich eine nummerische Unterstützung zum Veranschaulichen von Äquidistanz methodisch einwandfrei im Rahmen der Evaluation von Lehrveranstaltungen umsetzen lässt. Folgte man beispielsweise der Fachliteratur, sollten Antwortkategorien sprachlich von negativ zu positiv formuliert werden, da dies dem natürlichen Lesefluss von Testpersonen entspricht (Mummendey/Grau, 2014: 84), was dazu führte, dass die erste Antwortkategorie (negativ) nummerisch mit einer höheren Zahl verknüpft werden müsste als die letzte Antwortkategorie (positiv), die der Bestnote entspräche. Zwar ließe sich auch die Bestnote mit der höchsten Zahl versehen, jedoch entspräche diese Darstellung nicht mehr dem bekannten Schulnotenprinzip von 1 bis 6, wonach die beste Leistung die nummerisch kleinste Zahl erhält. Beide Prinzipien – die Formulierung der Antwortkategorien von negativ zu positiv sowie ein aufsteigendes Schulnotenprinzip – sind im Rahmen der Evaluation von Lehrveranstaltungen nur schwer in Einklang zu bringen. In den beiden in Kapitel *4 Ergebnis* dargestellten Entwürfen des *EEsL* wird einerseits ein Antwortformat von negativ zu positiv ohne nummerische Unterstützung (Tabelle 3, Entwurf 1) und ein Antwortformat von positiv zu

negativ mit nummerischer Unterstützung (Tabelle 4, Entwurf 2) präsentiert. Eine eindeutige Empfehlung – für oder gegen eines der beiden Prinzipien – wird an dieser Stelle nicht ausgesprochen, denn beide Prinzipien sind methodisch geeignet.

4 Ergebnis

Tabelle 3 und Tabelle 4 veranschaulichen zwei Entwürfe des *EEsL*. Der zweite Entwurf (vergleiche Tabelle 4) unterscheidet sich unter anderem dahingehend von dem ersten Entwurf als versucht wurde, den suggestiven Anteil der Fragen zu reduzieren. Die 10 Fragen sind inhaltlich sechs Bereichen zugeordnet, die aus Schmidts und Loßnitzers (2010: 58) Übersicht zu Gestaltungsmerkmalen lehrbezogenen Feedbacks übernommen wurden. Von den 13 möglichen Qualitätsdimensionen aus Schmidt und Loßnitzer (2010: 58) wurden nur sechs Dimensionen inhaltlich berücksichtigt, da der *EEsL* möglichst ökonomisch sein sollte und inhaltliche Redundanzen vermieden werden sollten. Die Qualitätsdimension *Rahmenbedingungen* wurde beispielsweise nicht operationalisiert, da diese Dimension auch über die offenen Fragen 9 und 10 erfasst werden kann. Insbesondere bei schlechten *Rahmenbedingungen* einer Lehrveranstaltung, beispielsweise einer zu geringen Raumkapazität, ist mit Freitextnennungen zu rechnen. Da aber beispielsweise die Raumkapazität bekannt und auch durch einen Lehrenden beobachtbar ist, muss sie nicht zwingend als einzelne Frage aufgenommen werden. Die Qualitätsdimension *Kompetenzerwerb* wurde nicht berücksichtigt, da für deren Operationalisierung eine eigenständige Klasse von Fragebögen verfügbar ist und innerhalb eines ökonomischen Fragebogens nicht berücksichtigt werden kann. Der Fokus bei der Erstellung des *EEsL* lag auf Fragen, die zur Beurteilung von *Lehrqualität* zwingend notwendig sind. Fragen zum Studiengang oder des derzeitigen Semesters, in dem ein Studierender eingeschrieben ist, werden nicht empfohlen. Sie werden deshalb nicht empfohlen, um auch in Lehrveranstaltungen mit geringer Teilnehmerzahl Anonymität zu gewährleisten. Die Gefahr sozial erwünschter Antworten bei subjektiv bedrohter Anonymität wiegt an dieser Stelle schwerer als ein möglicher Informationsgewinn durch weitere Fragen. Ebenfalls nicht mit aufgenommen wurde eine Frage nach der Häufigkeit, mit der ein Studierender die Lehrveranstaltung besucht. Um eine verlässlichere Übersicht zu Teilnahmequoten zu gewinnen, wird den Lehrenden das händische Zählen zu Beginn einer Lehrveranstaltung empfohlen, denn manche Information lässt sich auch ohne Fragebogen erheben (Reissert, 1992: 21). Der Autor selbst hat mit Kollegen über mehrere Semester auf diese Weise verlässlich die Anwesenheitsquote einer Lehrveranstaltung des ersten Studienabschnitts ermittelt.

Mitaufgenommen in den *EEsL* wurden zwei offene Fragen, sodass auch qualitative Auswertungen möglich sind. Diese Möglichkeit bot bereits der bestehende Fragebogen der Abteilung Betriebswirtschaft. Mitunter kann die Auswertung der Antworten auf offene Fragen sogar hilfreicher sein als die Auswertungen geschlossener Fragen (Schmidt/Loßnitzer, 2010: 68; Souvignier/Gold, 2003: 142), weshalb die beiden offenen Fragen die explizite Aufforderung enthalten, jeweils drei Nennungen zu geben. Ein mögliches Kategoriensystem zur Klassifikation von freitextlichen Nennungen geben beispielsweise Gediga et al. (2000: 59; 89-92).

Tabelle 3: Entwurf 1 – Evaluationsfragebogen zur Erfassung studentischer Lehrurteile (EEsL) (eigene Darstellung).

		stimme nicht zu	stimme eher nicht zu	teils/ teils	stimme eher zu	stimme zu	keine An-gabe
	Struktur						
1	Die Lehrveranstaltung hat für mich eine klar erkennbare Struktur (roter Faden).						
	Beitrag des Dozenten/der Dozentin						
2	Der Dozent/die Dozentin erklärt gut nachvollziehbar.						
3	Der Dozent/die Dozentin hat einen abwechslungsreichen Vortragsstil.						
4	Ich hatte Gelegenheit mich aktiv zu beteiligen.						
	Beitrag der Studierenden						
5	Die Studierenden haben aktiv zum Erfolg der Lehrveranstaltung beigetragen.						
	Interaktion						
6	In der Lehrveranstaltung herrscht eine störungsfreie Arbeitsatmosphäre.						
7	Es besteht ein angenehmes Klima zwischen Studierenden und Dozent/Dozentin.						
	Transformation						
8	Ich wurde zu eigenständigem Denken angeregt.						
	Qualitätsentwicklung (offene Fragen)						
9	Nennen Sie bitte drei Aspekte, die Ihnen gefallen haben.						
10	Nennen Sie bitte drei Aspekte, die Ihnen nicht gefallen haben.						

Tabelle 4 beinhaltet einen zweiten Entwurf des *EEsL*, der inhaltlich auf dem ersten Entwurf basiert. Im Unterschied zum ersten Entwurf (Tabelle 3) stellt der zweite Entwurf den Versuch dar, die Gefahr der sozialen Erwünschtheit in den Antworten der Testpersonen zu reduzieren. Am Beispiel des zweiten Items sei dieser Unterschied verdeutlicht. Während im ersten Entwurf (Tabelle 3) implizit eine Art positive Tendenz anhand der Wörter „gut nachvollziehbar" vorgegeben ist, wird im zweiten Entwurf (Tabelle 4) auf diese implizite Botschaft verzichtet. Im zweiten Entwurf (Tabelle 4) muss ein Studierender ohne implizite Botschaften zu einem Urteil gelangen, was auch anhand der veränderten Bezeichnungen der Antwortkategorien von *sehr gut* bis *sehr schlecht* ersichtlich wird. Zusätzlich sind die Antwortkategorien von positiv zu negativ formuliert und umfassen eine nummerische Unterstützung zum Verdeutlichen ihrer Äquidistanz. Im ersten Entwurf (Tabelle 3) wurde aufgrund der Formulierung von negativ zu

positiv auf eine nummerische Unterstützung verzichtet, weil sie sich methodisch nicht problemlos implementieren lässt. Zusätzlich wurde im zweiten Entwurf des *EEsL* die Paarform „Dozent/Dozentin“ – wodurch die Vielfalt der Geschlechter betont wird – in den Items durch den noch neutraleren Begriff „Lehrperson“ ersetzt, da an Hochschulen mitunter Empfehlungen hinsichtlich einer genderneutralen Sprache vorliegen und berücksichtigt werden müssen (vergleiche beispielsweise Frauenbeauftragte LMU München, 2011; Gäckle, 2015; Zentrale Frauenbeauftrage FU Berlin, 2016).

Tabelle 4: Entwurf 2 – Evaluationsfragebogen zur Erfassung studentischer Lehrurteile (EEsL) (eigene Darstellung).

		sehr gut	eher gut	teils/ teils	eher schlecht	sehr schlecht		keine An-gabe
		1	2	3	4	5		
	Struktur							
1	Die Struktur (roter Faden) der Lehrveranstaltung ist …							
	Beitrag der Lehrperson							
2	Die Lehrperson erklärt …							
3	Der Vortragsstil der Lehrperson ist …							
4	Die Gelegenheit mich aktiv zu beteiligen ist … vorhanden.							
	Beitrag der Studierenden							
5	Die aktive Beteiligung der Studierenden am Erfolg der Lehrveranstaltung ist …							
	Interaktion							
6	Die Arbeitsatmosphäre in der Lehrveranstaltung ist …							
7	Das Klima zwischen Studierenden und Lehrperson ist …							
	Transformation							
8	Die Möglichkeit zu eigenständigem Denken wird … gefördert.							
	Qualitätsentwicklung (offene Fragen)							
9	Nennen Sie bitte drei Aspekte, die Ihnen gefallen haben.							
10	Nennen Sie bitte drei Aspekte, die Ihnen nicht gefallen haben.							

Im letzten Kapitel *5 Diskussion* wird ein Fazit gezogen und ein Ausblick auf einen Einsatz des *EEsL* gegeben.

5 Diskussion

Großmann und Wolbring (2016: 8-9) nennen insgesamt sieben idealtypische Herausforderungen bei der Evaluation von Lehrveranstaltungen, die ihrer Ansicht nach zentral sind und anhand derer der *EEsL* beurteilt wird. Tabelle 5 veranschaulicht diese sieben idealtypischen Herausforderungen an eine Evaluation von Lehrveranstaltungen. Aufgrund ihres idealtypischen Charakters sind nicht immer alle Herausforderungen gleichermaßen in Einklang zu bringen, beispielsweise ein möglichst *umfassender Informationswert* versus eine möglichst hohe *Ökonomie*.

Tabelle 5: Idealtypische Herausforderungen der Evaluation von Lehrveranstaltungen (eigene Darstellung nach Großmann/Wolbring, 2016: 8-9).

idealtypische Herausforderungen	*Beurteilung*
möglichst umfassender Informationswert	steht im Widerspruch zur Ökonomie, die höher gewichtet wurde
Ökonomie	erfüllt
Partizipation und Akzeptanz der Stakeholder	teilweise erfüllt
Belastbarkeit der Ergebnisse	derzeit noch ungeklärt
Entscheidungsfunktion	derzeit noch nicht empfohlen
Legitimation für Entscheidungen	derzeit noch ungeklärt
Auswirkungen/Effektivität auf Lehre und Studium	derzeit noch ungeklärt

Erfüllt sind derzeit die Punkte *Ökonomie* sowie teilweise die *Partizipation und Akzeptanz der Stakeholder*. Stakeholder waren einerseits im Rahmen der Studienkommission beteiligt, andererseits durch den Einbezug von Vertretern des Fachschaftsrats Wirtschaft sowie weiteren studentischen Experten. Da die Vertreter des Fachschaftsrats Wirtschaft und die weiteren studentischen Experten in die Analyse des derzeit verwendeten Fragebogens mit einbezogen wurden, konnten auch die von Studierenden als relevant erachteten Aspekte in der Konzeption des *EEsL* berücksichtigt werden. Nichtsdestotrotz wurden bislang noch nicht alle verfügbaren Stakeholder in den Prozess einbezogen werden, weshalb der Punkt der *Partizipation und Akzeptanz* nur teilweise erfüllt ist. Weitere Herausforderungen der Evaluation von Lehrveranstaltungen sind derzeit noch nicht erfüllt, beziehungsweise ungeklärt. Im Rahmen der vorliegenden Arbeit war es allerdings auch nicht das Ziel, die *Lehrqualität* mit einer solchen Präzision zu erfassen, sodass die Ergebnisse beispielsweise die Grundlage für Ressourcenzuteilungen oder Personalentscheidungen bilden können. Dies hätte einen deutlich umfangreicheren Fragebogen erfordert. Vielmehr sollte im Rahmen der vorliegenden Arbeit ein ökonomischer und methodisch fundierter Fragebogen zur Evaluation von Lehrveranstaltungen vorgeschlagen werden, der einerseits den Kommunikationsprozess in der Abteilung Betriebswirtschaft zu Lehre angemessen unterstützt, andererseits gesetzliche Vorgaben erfüllt. Der *EEsL* erfüllt das Gütekriterium der *Ökonomie* besser als der derzeit verwendete Fragebogen zur Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft, da er nur rund ein Drittel der bisherigen Anzahl an Fragen umfasst. Die Kürze des *EEsL* macht ihn zwar dahingehend kritisierbar, als er allgemein gehalten ist und einem Lehrenden wenig konkrete Anregungen bietet (Tillmann et al., 2011: 81), dieser Nachteil erscheint aufgrund der empirischen Redundanz theoretisch-inhaltlich unabhängiger Dimensionen des derzeit verwendeten Fragebogens aber wenig erheblich. Inwieweit Studierende unter Verwendung des *EEsL* inhaltlich zwischen theoretisch unabhängigen Dimensionen unterscheiden, sollte empirisch überprüft werden. Ungeklärt ist derzeit noch, ob die hohen Zusammenhänge theoretisch-inhaltlich unabhängiger Dimensionen aufgrund metho-

discher Mängel des derzeit verwendeten Fragebogens resultieren oder einer zu geringen Differenzierungsleistung der Studierenden beim Ausfüllen geschuldet sind. Der *EEsL* erfüllt methodische Richtlinien. Verglichen mit dem derzeit verwendeten Fragebogen wurden doppeldeutige Items vermieden, nur Items berücksichtigt, die inhaltlich sinnvoll zu beurteilen sind, und auf variierende, unterschiedlich beschriftete Antwortkategorien verzichtet. Zudem wurden Empfehlungen hinsichtlich der Anzahl und der Beschriftung von Antwortkategorien sowie zur Durchführung einer Evaluation von Lehrveranstaltungen gegeben. Inwieweit Entwurf 1 oder Entwurf 2 des *EEsL* jeweils vorzuziehen ist, sollte von den Entscheidungsträgern vor der Implementierung diskutiert werden. Beide Entwürfe haben ihre eigenen Stärken sowie Schwächen. Zwar stellt Entwurf 2 (Tabelle 4) des *EEsL* den Versuch dar, sozial erwünschte Antworten zu reduzieren, gleichzeitig ist aber die implizit gesendete Botschaft dessen, was unter guter Lehre verstanden wird, geringer. Ähnliches gilt für die Reihenfolge der Antwortkategorien. Es finden sich gute Gründe für eine Reihenfolge von positiv zu negativ mit nummerischer Unterstützung, gleichzeitig finden sich aber auch gute Gründe für eine Reihenfolge von negativ zu positiv ohne nummerische Unterstützung. Auch lässt sich trefflich darüber debattieren, ob eine separate Antwortkategorie „keine Angabe“ im Rahmen der Evaluation von Lehrveranstaltungen notwendig ist oder nicht. Inwieweit anstelle der methodisch eher unglücklichen Paarform „Dozent/Dozentin“ ein Begriff wie beispielsweise „Lehrperson“ verwendet werden sollte, ist ebenfalls eine Geschmacksfrage. Zwar führt das Wort „Lehrperson“ zu kürzeren und damit prägnanteren Items, gleichzeitig ist es aber alltagssprachlich weniger repräsentiert. Festzuhalten ist, dass beide Entwürfe einen Kern an Items umfassen, der systematisch hergeleitet wurde und verwendet werden kann. Hinsichtlich der methodischen Rahmung der Items besteht jedoch Spielraum.

Unberücksichtigt geblieben sind im *EEsL* Verzerrungsvariablen, die aber auch nachträglich noch aufgenommen werden können. Konzentriert man sich dabei lediglich auf einen Kern an Verzerrungsvariablen (siehe *2.3 Verzerrungsvariablen*) ist auch weiterhin ein ökonomischer Fragebogen möglich. Aus Sicht des Autors ist die Aufnahme von möglichen Verzerrungsvariablen aber nur dann unerlässlich, wenn anhand der Daten auch inferenzstatistische Zusammenhänge zurückgemeldet werden und nicht nur deskriptive wie bislang. Ein Nebenprodukt dieser Arbeit ist, dass der *EEsL* auch zur Verwendung in der beruflichen Praxis geeignet ist. Keines der Items umfasst spezifische Formulierungen, die eine Verwendung lediglich auf den Hochschulbereich begrenzen. Die Bezeichnung „Studierende“ lässt sich in der beruflichen Praxis beispielsweise durch „Teilnehmende“ ersetzen.

Inwieweit der *EEsL* – abgesehen von den Gütekriterien *Ökonomie* und *Objektivität* – die Gütekriterien *Reliabilität* und *Validität* erfüllt, kann zum aktuellen Zeitpunkt noch nicht beurteilt werden. Um beispielsweise erste Reliabilitätseinschätzungen vornehmen zu können, wird ein Einsatz im Rahmen der Evaluation von Lehrveranstaltungen in der Abteilung Betriebswirtschaft empfohlen. Aufwändigere Messungen zur Einschätzung der Validität sollten jedoch frühestens in einem zweiten Schritt durchgeführt werden und sind derzeit noch nicht empfehlenswert. Nun sollten zunächst Daten mit dem *EEsL* gesammelt und über weitere Schritte datenbasiert entschieden werden.

6 Literatur

Alvensleben, B./Morsch, R./Schirmer, J. (1978). Veranstaltungsbegleitende Kritik unter ständiger Mitwirkung einer studentischen Arbeitsgemeinschaft (S. 141-158). In: L. Huber/I. Bürmann/R. Francke/W. Schmidt (Hrsg.). *Auswertung. Rückmeldung. Kritik im Hochschulunterricht. Band I: Einführung und Überblick*. Hamburg: Arbeitsgemeinschaft für Hochschuldidaktik e.V.

Antretter, E./Dunkel, D./Haring, C. (2013). Wie zeitgemäß sind die in der deutschsprachigen Erwachsenenpsychiatrie verwendeten psychologischen Leistungstests? Eine Übersichtsarbeit. *Psychiatrische Praxis*, 40 (3), S. 120-129.

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

Berger, U./Schleußner, C./Strauß, B. (2003). Umfassende Lehrevaluation in der Medizin – eine Aufgabe für die psychosozialen Fächer? *Psychotherapie, Psychosomatik, medizinische Psychologie*, 53 (2), S. 71-78.

Blossfeld, H.-P. (2010). Survival- und Ereignisanalyse (S. 995-1016). In: C. Wolf/H. Best (Hrsg.). *Handbuch der sozialwissenschaftlichen Datenanalyse*. Wiesbaden: Verlag für Sozialwissenschaften.

Böttcher, W./Grewe, C.M. (2010). Eine Untersuchung zur Wirksamkeit der studentischen Lehrveranstaltungskritik am Beispiel der Westfälischen Wilhelms-Universität Münster (S. 73-82). In: P. Pohlenz/A. Oppermann (Hrsg.). *Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation?* Bielefeld: UVW.

Braun, E. (2007). Ergebnisorientierte Lehrveranstaltungsevaluation: Das Berliner Evaluationsinstrument für studentische Kompetenzen (S. 73-82). In: A. Kluge/K. Schüler (Hrsg.). *Qualitätssicherung und –entwicklung in der Hochschule: Methoden und Ergebnisse*. Lengerich: Pabst Science Publishers.

Braun, E. (2008). *Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen (BEvaKomp)*. Göttingen: V&R unipress.

Braun, E./Gusy, B. (2006). Perspektiven der Lehrevaluation (S. 152-166). In: G. Krampen/H. Zayer (Hrsg.). *Didaktik und Evaluation in der Psychologie*. Göttingen: Hogrefe.

Braun, E./Gusy, B./Leidner, B./Hannover, B. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica*, 54 (1), S. 30-42.

Brehl, A./Schneider, D./Utsch, A. (2004). Studentische Lehrevaluation mit EvaSys am Beispiel der FH Nordostniedersachsen. *Zeitschrift für Evaluation*, 2/2004, S. 311-323.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Auflage). Hallbergmoos: Pearson.

Bundesministerium für Bildung und Forschung (BMBF) (2016). *Bildung und Forschung in Zahlen 2016. Ausgewählte Fakten aus dem Daten-Portal des BMBF. www.datenportal.bmbf.de*. Berlin: Bundesministerium für Bildung und Forschung.

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

Diehl, J.M. (2002). *VBVOR – VBREF. Fragebögen zur studentischen Lehrevaluation von Hochschulveranstaltungen. Manual*. Gießen: Justus-Liebig-Universität Gießen.

Diehl, J.M. (2003). Normierung zweier Fragebögen zur studentischen Beurteilung von Vorlesungen und Seminaren. *Psychologie in Erziehung und Wissenschaft*, 50 (1), S. 27-42.

Döring, N./Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Auflage). Berlin: Springer.

Dresel, M./Tinsner, K. (2008). Onlineevaluationen von Lehrveranstaltungen: Methodeneffekte bei der Onlineevaluation von Lehrveranstaltungen. *Zeitschrift für Evaluation*, 7 (2), S. 183-211.

el Hage, N. (1996). *Lehrevaluation und studentische Veranstaltungskritik. Projekte, Instrumente und Grundlagen*. Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF).

Europäische Bildungsminister (1999*). The Bologna Declaration of 19 June 1999. Joint Declaration of the European Ministers of Education*. [Elektronische Ressource]. Online verfügbar unter: http://media.ehea.info/file/Ministerial_conferences/02/8/1999_Bologna_Declaration_English_553028.pdf [26.09.2016].

Fiedler, B./Billmann-Mahecha, E. (1997). *Entwicklung und Erprobung eines Fragebogens zur Evaluation von Lehrveranstaltungen durch Studierende*. Institut für Psychologie und Soziologie in den Erziehungswissenschaften. Abteilung Psychologie. Forschungsbericht Nr. 1/1997. Hannover: Universität Hannover.

Fischer, V. (2014). Die Evaluation von Lehrveranstaltungen an der Medizinischen Hochschule Hannover. *Qualität in der Wissenschaft*, 8 (2/3), S. 41-46.

Fischer, V. (2016). *Kurzes Inventar zur Evaluation durch Studierende (KIES)*. Persönliche Mitteilung via E-Mail, empfangen am 18.08.2016.

Fondel, E./Lischetzke, T./Weis, S./Gollwitzer, M. (2014). Zur Validität von studentischen Lehrveranstaltungsevaluationen. Messinvarianz über Veranstaltungsarten, Konsistenz von Urteilen und Erklärung ihrer Heterogenität. *Diagnostica*, 61 (3), S. 124-135.

Frauenbeauftragte der Ludwig-Maximilians-Universität München (2011). *Leitfaden gendergerechte Sprache*. [Elektronische Ressource]. Online verfügbar unter: http://www.frauenbeauftragte.uni-muenchen.de/genderkompetenz/sprache/index.html [07.12.2016].

Gäckle, A. (2015). *ÜberzeuGENDERe Sprache. Leitfaden für eine geschlechtersensible und inklusive Sprache*. [Elektronische Ressource]. Online verfügbar unter: http://www.gb.uni-koeln.de/e2106/e2113/e5726/2014_Leitfaden_UeberzeuGENDEReSprache_11032014.pdf [07.12.2016].

Gaens, T. (2015). Noteninflation an deutschen Hochschulen – Werden die Examensnoten überall immer besser? *Beiträge zur Hochschulforschung*, 37 (4), S. 8-35.

Gediga, G., Kannen, K. v., Schnieder, F., Kohne, S., Luck, H. & Schneider, B. (2000). *Kiel: Ein Kommunikations-Instrument für die Evaluation von Lehrveranstaltungen. Bericht über die Entwicklung und Anwendungsmöglichkeiten einer fonnativen Evaluationsprozedur im universitären Alltag.* Bangor: Methodos.

Gläßer, E./Gollwitzer, M./Kranz, D./Meiniger, C./Schlotz, W./Schnell, T./Voß, A. (2002). Das „Trierer Inventar zur Lehrveranstaltungsevaluation“. [Elektronische Ressource]. Online verfügbar unter: https://www.zpid.de/pub/tests/4523_Fragebogen_TRIL3_w.pdf [07.05.2016].

Gollwitzer, M./Schlotz, W. (2003). Das „Trierer Inventar zur Lehrveranstaltungsevaluation“ (TRIL): Entwicklung und erste testtheoretische Erprobungen (S. 114-128). In: G. Krampen/H. Zayer (Hrsg.). *Psychologiedidaktik und Evaluation IV*. Bonn: Deutscher Psychologen Verlag.

Gollwitzer, M./Kranz, D./Vogel, E. (2006). Die Validität studentischer Lehrveranstaltungsevaluationen und ihre Nützlichkeit für die Verbesserung der Hochschullehre: Neuere Befunde zu den Gütekriterien des „Trierer Inventars zur Lehrevaluation“ (TRIL) (S. 90-104). In: G. Krampen/H. Zayer (Hrsg.). *Didaktik und Evaluation in der Psychologie*. Göttingen: Hogrefe.

Greenwald, G. (2013). *NSA collecting phone records of millions of Verizon customers daily*. [Elektronische Ressource]. Online verfügbar unter: https://www.theguardian.com/world/2013/jun/06 /nsa-phone-records-verizon-court-order [08.08.2016].

Greenwald, G./MacAskill, E./Poitras, L. (2013). *Edward Snowden: the whistleblower behind the NSA surveillance revelations*. [Elektronische Ressource]. Online verfügbar unter: https://www.theguardian .com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance [08.08.2016].

Großmann, D./Wolbring, T. (2016). Stand und Herausforderungen der Evaluation an deutschen Hochschulen (S. 3-25). In: D. Großmann/T. Wolbring (Hrsg.). *Evaluation von Studium und Lehre. Grundlagen, methodische Herausforderungen und Lösungsansätze*. Wiesbaden: Springer.

Guilford, J.P. (1954). *Psychometric Methods* (2. Auflage). New York: McGraw Hill.

Gustad, J.W. (1967). Evaluation of Teaching Performance: Issues and Possibilities (S. 265-281). In: C.B.T. Lee (Hrsg.). *Improving College Teaching*. Washington, D.C.: American Council on Education.

Hansen, U./Hennig-Thurau, T./Wochnowski, H. (1997). TEACH-Q: Ein valides und handhabbares Instrument zur Bewertung von Vorlesungen. *Die Betriebswirtschaft (DBW)*, 57 (3), S. 376-396.

Heber, F. (2013). *Unveröffentlichter Datensatz*. Hannover: Hochschule Hannover.

Hochschule Hannover (2006). *Ordnung zur internen Lehrevaluation. Verkündungsblatt der FHH Nr. 3/2006 vom 1.3.2006*. [Elektronische Ressource]. Online verfügbar unter: http://www.hs-hannover.de/fileadmin/media/doc/pp/verkuendungsblatt/Ordnung_zur_internen_Lehrevaluation.pdf [30.09.2016]. Hannover: Hochschule Hannover.

Höft, S. (2014). Erfolgsüberprüfung personalpsychologischer Arbeit (S. 1081-1135). In: H. Schuler/U.P. Kanning (Hrsg.). *Lehrbuch der Personalpsychologie* (3. Auflage). Göttingen: Hogrefe.

Hofmann, J.M. (1990). Die Beurteilung pädagogisch-psychologischer Lehrveranstaltungen anhand des VP-Psych. *Psychologie in Erziehung und Unterricht*, 37, S. 47-53.

Jonkisz, E./Moosbrugger, H./Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen (S. 27-74). In: H. Moosbrugger/A. Kelava (Hrsg.). *Testtheorie und Fragebogenkonstruktion* (2. Auflage). Berlin: Springer.

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

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

Krawietz, M. (2006). *HISBUS-Kurzinformation Nr. 16. Evaluation der Evaluationen. Erfolg und Misserfolg von Evaluationen an deutschen Hochschulen – Die Perspektive der Studierenden*. Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.

Kubinger, K.D./Proyer, R. (2010). Gütekriterien (S. 173-180). In: K. Westhoff/C. Hagemeister/M. Kersting/ F. Lang/H. Moosbrugger/G. Reimann/G. Stemmler (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Auflage). Lengerich: Pabst Science.

Landes, T./Ziegler, M. (2015). Ein Praxisbeispiel zur Konstruktion eines Lehrevaluationsinstruments: Berliner Lehrevaluationsinventar für Vorlesungen (BLEI-VL). *Diagnostica*, 61 (3), S. 136-143.

Landesamt für Statistik Niedersachsen (2016). *Deutsche und ausländische Studierende und Studienanfänger/-innen im Sommersemester 2016 nach Hochschulart und Hochschule – vorläufiges Ergebnis – Korrigierte Fassung vom 18.08.2016*. [Elektronische Ressource].

Online verfügbar unter: http://www.statistik.niedersachsen.de/download/108715 [30.09.2016]. Hannover: Landesamt für Statistik Niedersachsen.
Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest. Manual zum MWT-B* (5. Auflage). Balingen: Spitta.
Linssen, R. (2016). Leserbrief. *Die Neue Hochschule*, 5, S. 156.
Linssen, R./Mayer, M. (2016). „Sprache ist die Basis der Grundlage des Fundaments …". Zu Sprach- und Lesekompetenzen von Studierenden. *Die Neue Hochschule*, 2, S. 42-45.
Litzcke, S. (2015). *Studentische Lehrevaluation im Wintersemester 2014/2015 und im Sommersemester 2015 in den Studiengängen der Abteilung Betriebswirtschaft*. Interner Bericht. Hannover: Hochschule Hannover.
Litzcke, S./Linssen, R./Maffenbeier, S./Schilling, J. (2012). *Korruption: Risikofaktor Mensch. Wahrnehmung – Rechtfertigung – Meldeverhalten*. Wiesbaden: Springer.
Lohnert, B./Rolfes, M. (1997). *Handbuch zur Evaluation von Lehre und Studium an Hochschulen. Ein praxisorientierter Leitfaden*. Hannover: Zentrale Evaluationsagentur der niedersächsischen Hochschulen.
Loßnitzer, T./Schmidt, B./Born, S. (2007). Zentrale Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Qualitätsmodell und Messinstrument (S. 327-335). In: M. Krämer/S. Preiser/K. Brusdeylins (Hrsg.). *Psychologiedidaktik und Evaluation VI. Band 9*. Göttingen: V&R unipress.
Lozar Manfreda, K./Bosnjak, M./Berzelak, J./Haas, I./Vehovar, V. (2008). Web surveys versus other survey modes. A meta-analysis comparing response rates. *International Journal of Market Research*, 50 (1), S. 79-104.
Lück, S. (2005). *Studentische Veranstaltungskritik: Eine kritische Würdigung der statistischen Anforderungen*. Aachen: Shaker.
Meinefeld, W. (2010). Online-Befragungen im Kontext von Lehrevaluationen – praktisch und unzuverlässig. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 62 (2), S. 297-315.
Metje, B./Kelle, U. (2010). Qualitätsentwicklung von Lehrveranstaltungsevaluationen durch Methodenkombination (S. 97-107). In: P. Pohlenz/A. Oppermann (Hrsg.). *Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation?* Bielefeld: UVW.
Moosbrugger, H./Krömker, D. (2011). *STUD-FEL Studierenden-Fragebogen zur Evaluation von Lehrveranstaltungen V 1.5*. [Elektronische Ressource]. Online verfügbar unter: https://www.uni-frankfurt.de/37364975/lehreval_fragebogen_ohne-la.pdf [29.08.2016].
Müller-Wolf, H.-M. (1977). *Lehrverhalten an der Hochschule. Dimensionen, Zusammenhänge, Trainingsmöglichkeiten.* München: Verlag Dokumentation.
Mummendey, H.D./Grau, I. (2014). *Die Fragebogen-Methode. Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung* (6. Auflage). Göttingen: Hogrefe.
Paechter, M./Maier, B./Macher, D. (2011). Evaluation universitärer Lehre mittels Einschätzungen des subjektiven Kompetenzerwerbs. *Psychologie in Erziehung und Unterricht*, 58 (2), S. 128-138.
Peter, J./Leichner, N./Mayer, A.-K./Krampen, G. (2014a). Das Inventar zur Evaluation von Blended Learning (IEBL): Konstruktion und Erprobung in einem Training professioneller Informationskompetenz (S. 275-282). In: M. Kraemer/U. Weger/M. Zupanic (Hrsg.). *Psychologiedidaktik und Evaluation X*. Aachen: Shake.
Peter, J./Leichner, N./Mayer, A.-K./Krampen, G. (2014b). *Das Inventar zur Evaluation von Blended Learning (IEBL): Konstruktion und Erprobung in einem Training professioneller*

Informationskompetenz. Vortrag auf 10. Fachtagung Psychologiedidaktik und Evaluation. 30. bis 31.05.14, Witten.

Peters, B./Reiß, S./Tillmann, A./Schweizer, K. (2011). Untersuchungen zur psychometrischen Qualität des Frankfurter Lehrveranstaltungsfragebogens mit dem Bifaktorenmodell. In: M. Krämer/S. Preiser/K. Brusdeylins (Hrsg.). *Psychologiedidaktik und Evaluation VIII*. Aachen: Shaker.

Pötschke, M. (2009). Potentiale von Online-Befragungen: Erfahrungen aus der Hochschulforschung (S. 74-89). In: N. Jackob/H. Schoen/T. Zerback (Hrsg.). *Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Pohlenz, P./Oppermann, A. (2010). Wie viel Wissenschaft braucht die Evaluation? Eine Einführung. (S. 3-16). In: P. Pohlenz/A. Oppermann (Hrsg.). *Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation?* Bielefeld: UVW.

Reischmann, J. (1995). Kursbeurteilungsbogen KBB: Ein Fragebogeninstrument zur Messung der Qualität von Weiterbildungskursen (S. 269-279). In: R. Arbinger/R.S. Jäger (Hrsg.). *Zukunftsperspektiven empirisch-pädagogischer Forschung (Empirische Pädagogik, Beiheft 4)*. Landau: Empirische Pädagogik.

Reissert, R. (1992). Dokumentation. Evaluation der Lehre Teil 1. Aktuelle Aktivitäten an deutschen Hochschulen. Hannover: Hochschul-Informations-System (HIS).

Rieck, W. (1978). Teilnehmerorientierte Unterrichtskritik als Mittel der Weiterentwicklung und Neuplanung einer regelmäßig angebotenen Lehrveranstaltung (S. 213-224). In: L. Huber/I. Bürmann/R. Francke/W. Schmidt (Hrsg.). *Auswertung. Rückmeldung. Kritik im Hochschulunterricht. Band I: Einführung und Überblick*. Hamburg: Arbeitsgemeinschaft für Hochschuldidaktik e.V.

Rindermann, H. (2009). *Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (2. Auflage). Landau: Verlag Empirische Pädagogik.

Rindermann, H. (2016). Lehrveranstaltungsevaluation an Hochschulen (S. 227-262). In: D. Großmann/T. Wolbring (Hrsg.). *Evaluation von Studium und Lehre. Grundlagen, methodische Herausforderungen und Lösungsansätze*. Wiesbaden: Springer.

Rindermann, H./Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE). Handanweisung*. Heidelberg: Roland Asanger.

Rindermann, H./Baumeister, A.E.E./Gröper, A. (2014). Cognitive abilities of Emirati and German engineering university students. *Journal of Biosocial Science*, 46 (2). S. 199-213.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 9 (3), S. 222-245.

Rost, J. (2004). *Lehrbuch. Testtheorie – Testkonstruktion* (2. Auflage). Bern: Hans Huber.

Saris, W.E./Gallhofer, I.N. (2007). *Design, Evaluation, and analysis of questionnaires for survey research*. Hoboken, New Jersey: John Wiley & Sons.

Satzger, W./Fessmann, H./Engel, R.R. (2002). Liefern HAWIE-R, WST und MWT-B vergleichbare IQ-Werte? *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23 (2), S. 159-170.

Schafferer, W.G. (2010). *Lehrevaluation und Studiensituation an der Universität Innsbruck. Bologna Prozess – Universitätsgesetz – Qualitätsmanagement*. Frankfurt am Main: Peter Lang.

Schmidt, B./Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of the Art, ein Definitionsvorschlag und Entwicklungslinien. *Zeitschrift für Evaluation*, 9 (1), S. 49-72.

Schön, F. (2011). *Korruption. Wie eine Hand die andere wäscht*. Frankfurt: Verlag für Polizeiwissenschaft.

Schone, R. (2016). Leserbriefe zum Aufsatz „Sprache ist die Basis der Grundlage des Fundaments …“ Zu Sprach- und Lesekompetenzen von Studierenden von Ruth Linssen und Maike Meyer (DNH Heft 2/2016). *Die Neue Hochschule*, 4, S. 122.

Schroer, J. (2003). *Studentische Lehrveranstaltungsevaluation im Internet. Trier: Universität Trier*. [Elektronische Ressource]. Online verfügbar unter: https://www.uni-trier.de/fileadmin/fb1/ein/PLA/Schroer__J.__2003_._Studentische_Lehrveranstaltungsevaluation_im_Internet.pdf [05.08.2016].

Schuler, H. (2014). 3. Arbeits- und Anforderungsanalyse (S. 61-98). In: H. Schuler/U.P. Kanning (Hrsg.). *Lehrbuch der Personalpsychologie* (3. Auflage). Göttingen: Hogrefe.

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

Shih, T.-H./Fan, X. (2008). Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. *Field Methods*, 20 (3), S. 249-271.

Simon, A./Zajontz, Y./Reit, V. (2013). Lehrevaluation online oder papierbasiert? Ein empirischer Vergleich zwischen traditionellem Fragebogen und inhaltsgleicher Online-Erhebung. *Beiträge zur Hochschulforschung*, 35 (3), S. 8-26

Simonson, J./Pötschke, M. (2006). Akzeptanz internetgestützter Evaluationen an Universitäten. *Zeitschrift für Evaluation*, 5 (2), S. 227-248.

Sommer, J. (1978). Ein Fragebogen zur Unterrichtssituation mit deskriptiven und präskriptiven Urteilen (S. 170-184). In: L. Huber/I. Bürmann/R. Francke/W. Schmidt (Hrsg.). *Auswertung. Rückmeldung. Kritik im Hochschulunterricht. Band I: Einführung und Überblick.* Hamburg: Arbeitsgemeinschaft für Hochschuldidaktik e.V.

Souvignier, E./Gold, A. (2003). Lehrevaluation als Feedback für Lehrende: Entwicklung eines Fragebogens unter hochschuldidaktischer Perspektive (S. 129-144). In G. Krampen/H. Zayer (Hrsg.). *Psychologiedidaktik und Evaluation IV*. Bonn: Deutscher Psychologen Verlag.

Statistisches Bundesamt (2013). *Berufliche Weiterbildung in Unternehmen. Vierte europäische Erhebung über die berufliche Weiterbildung in Unternehmen (CVTS4)*. Wiesbaden: Statistisches Bundesamt.

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

Stern, E./Grabner, R.H. (2014). Die Erforschung menschlicher Intelligenz (S. 174-201). In: L. Ahnert (Hrsg.). *Theorien in der Entwicklungspsychologie*. Berlin: Springer.

Stratmann, F. (2016). § 5 Evaluation von Forschung und Lehre (S. 144-154). In: V. Epping (Hrsg.). *Niedersächsisches Hochschulgesetz mit Hochschulzulassungsgesetz. Handkommentar*. Baden-Baden: Nomos.

Seyda, S./Werner, D. (2014). IW-Weiterbildungserhebung 2014 – Höheres Engagement und mehr Investitionen in betriebliche Weiterbildung. *IW-Trends – Vierteljahresschrift zur empirischen Wirtschaftsforschung*, 41 (4), S. 1-15.

Thielsch, M.T./Hirschfeld, G. (2012). Münsteraner Fragebogen zur Evaluation von Vorlesungen – revidiert (MFE-Vr). In: A. Glöckner-Rist (Hrsg.). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. ZIS Version 15.0. Bonn: GESIS.

Thomas, A. (2001). Interkulturelle Kompetenz in der internationalen wissenschaftlichen Zusammenarbeit (S. 219-236). In: G. Fink/S. Meierewert (Hrsg.). *Interkulturelles Management: österreichische Perspektiven*. Wien: Springer.

Thomas, A. (2016). *Interkulturelle Psychologie. Verstehen und Handeln in internationalen Kontexten*. Göttingen: Hogrefe.

Tillmann, A./Reiß, S./Moosbrugger, H./Krömker, D./Schweizer, K./Gold, A. (2011). Qualitätssicherung der Lehre an großen Universitäten: Psychometrische Studien zum Frankfurter Studierendenfragebogen zur Evaluation von Lehrveranstaltungen (STUD-FEL). *Qualität in der Wissenschaft - Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration*, 5 (3), S. 79-89.

Tinsner, K./Dresel, M. (2007). Onlinebefragungen in der Lehrveranstaltungsevaluation: Ein faires, verzerrungsfreies und ökonomisches Verfahren (S. 59-69). In: A. Kluge/K. Schüler (Hrsg.). *Qualitätssicherung und –entwicklung in der Hochschule: Methoden und Ergebnisse*. Lengerich: Pabst Science Publishers.

Tröster, H./Gundlach, G./Moschner, B. (1997). Was erwarten Studierende der Psychologie von ihrer Diplomarbeit? *Zeitschrift für Pädagogische Psychologie*, 11 (2), S. 109-122.

van Randenborgh, A. (2016). Leserbriefe zum Aufsatz „Sprache ist die Basis der Grundlage des Fundaments …“ Zu Sprach- und Lesekompetenzen von Studierenden von Ruth Linssen und Maike Meyer (DNH Heft 2/2016). *Die Neue Hochschule*, 4, S. 123.

Webler, W.-D. (2005). Zur Bewertung von Lehrveranstaltungen. Konzeptionelle Begründung des Bielefelder Modells der Evaluation von Lehrveranstaltungen. *Das Hochschulwesen*, 53 (2), S. 63-70.

Webler, W.-D. (2010). Evaluation von Lehre und Studium als Hypothesenprüfung. (S. 33-53). In: P. Pohlenz/A. Oppermann (Hrsg.). *Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation?* Bielefeld: UVW.

Weis, S./Karthaus, C./Lischetzke, T. (2014). Elemente der Lehrveranstaltungsevaluation an der Universität Koblenz-Landau: Theoretische Einordung und empirische Befunde. *Qualität in der Wissenschaft*, 8 (2/3), S. 61-69.

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), S. 133-166.

Winteler, A./Schmolck, P. (1979). Entwicklung und Validierung eines Schätzverfahrens zur Beurteilung von Lehrveranstaltungen. *Schweizerische Zeitschrift für Psychologie*, 38 (2), S. 139-156.

Wolbring, T. (2013). *Fallstricke der Lehrevaluation. Möglichkeiten und Grenzen der Messbarkeit von Lehrqualität*. Frankfurt: Campus.

Wolf, P./Spiel, C./Pellert, A. (2001). Entwicklung eines Fragebogens zur globalen Lehrveranstaltungsevaluation – ein Balanceakt zwischen theoretischem Anspruch, Praktikabilität und Akzeptanz (S. 89-109). In: C. Spiel (Hrsg.). *Evaluierung an der Universität - zwischen Qualitätsmanagement und Selbstzweck*. Münster: Waxmann.

Zentrale Frauenbeauftrage der Freien Universität Berlin (2016). *Gender. Geschlechtersensible Sprache*. [Elektronische Ressource]. Online verfügbar unter: http://www.fu-berlin.de/sites/frauenbeauftragte/media/FU-Frauenbeauftragte-Flyer-2014-x30-web-geschlechtergerechtigkeit.pdf [07.12.2016].

Ziegler, M./Weis, S. (2015). Editorial. Lehrevaluation als Mittel zur Erfassung und Verbesserung universitärer Lehre? Methodische, diagnostische und inhaltliche Aspekte. *Diagnostica*, 61 (3), S. 113-115.

ZSW Hochschuldidaktik (ohne Jahr). *Beobachtung der Lernprozesse als Nachfolger der Lehrevaluation klassischen Stils*. [Elektronische Ressource]. Online verfügbar unter: http://www.hs-hannover.de/fileadmin/media/doc/hdidaktik/LEV_neu_Erlaeuterungen.pdf [30.09.2016]. Hannover: Hochschule Hannover.

Zumbach, J./Spinath, B./Schahn, J./Friedrich, M./Kögel, M. (2007). Entwicklung einer Kurzskala zur Lehrevaluation (S. 317-325). In: M. Krämer/S. Preiser/K. Brusdeylins (Hrsg.). *Psychodidaktik und Evaluation Band IV*. Göttingen: V&R unipress.