Nicole Burzan & Isa Jahnke (2010):
Was misst die studentische Lehrkritik? Eine empirische Infragestellung von Lehrevaluationen an Hochschulen.
In: SOZIOLOGIE, 39. Jg., Heft 4, 2010, S. 438-461.

SOZIOLOGIE, 39. JG., HEFT 4, 2010, S. XXX–XXX

Was misst die studentische Lehrkritik?

Eine empirische Infragestellung von Lehrevaluationen an Hochschulen

Nicole Burzan und Isa Jahnke

1. Einleitung: Evaluation oder Evaluitis?

Hochschulen und Universitäten befinden sich zurzeit in einem Identitätswandel (Würmseer 2010) und sind dem Druck ausgesetzt, in Forschung und Lehre konkurrenzfähig zu sein, z.B. um genügend Drittmittel, aber auch hinreichend viele Bachelor- und insbesondere Masterstudierende zu bekommen. Verschiedene Rankings und auch durch die Einrichtung von Exzellenz-Universitäten und Kompetenzzentren fördern diese Entwicklung. So befinden sich Hochschulen in kontinuierlichen Qualitätsoptimierungsprozessen. Interne Aufbau- und Ablauforganisationen stehen auf dem Prüfstand, um mit Universitäten mit Exzellenz-Status mithalten zu können oder besser zu werden. Diese Veränderungen betreffen Forschung, Selbstverwaltung und Lehre.

In diesem Beitrag fokussieren wir den Anspruch der Qualitätssicherung im Bereich der Lehre und stellen die *Evaluationspraxis* in Frage. Prinzipiell können Evaluationen als geeignete Instrumente zur Organisationsentwicklung und Qualitätssicherung angesehen werden. Zu inhaltlichen Argumenten für Evaluationen kommt ein rechtliches hinzu: Im Hochschulfreiheitsgesetz ist die Qualitätssicherung der Lehre durch Evaluationen festlegt, auf

deren Basis Maßnahmen zur Optimierung der Lehre abgeleitet werden sollen (s. Hochschulfreiheitsgesetz, HFG NRW v. 01.01.2007 §7 (2): »Zur Qualitätsentwicklung und -sicherung überprüfen und bewerten die Hochschulen regelmäßig die Erfüllung ihrer Aufgaben, insbesondere im Bereich der Lehre«; vgl. auch Pohlenz 2008: 68).

Allerdings zeigt sich gerade in den letzten Jahren eine aufkommende »Steuerungswut« (Ernst 2008) und teilweise wenig konzeptorientierte Datensammlung, die sich in einer als pathologisch kritisierten »Evaluitis« niederschlägt (Frey 2006; Döring 2006). So wird etwa moniert, dass »für eine Fülle von Merkmalen, mit denen sich Studiengänge oder auch einzelne Lehrveranstaltungen beschreiben lassen (…) bislang keine theoretisch und/ oder empirisch begründeten und konsensfähigen Qualitätsstandards entwickelt worden« sind (Döring 2006: 4). Diese Kritik dämpft den Steuerungsoptimismus durch das Wissen über potentielle Schwachstellen ebenso wie die in der Soziologie generell eher skeptisch betrachtete Diagnose einer »Wissensgesellschaft«. Der optimistischen Konnotation, durch ständige Wissenserweiterung Problemlösungen finden zu können, setzen diese kritischen Ansätze z.B. entgegen, dass mit dem Wissen auch das Nichtwissen zunimmt (Wehling 2007; vgl. auch Ammon et al. 2007; Bittlingmayer, Bauer 2006; Tänzler, Knoblauch 2006). Somit stellt sich die Frage, welchen Nutzen die Evaluation von Lehrveranstaltungen an Hochschulen für deren Qualitätsoptimierung erbringen kann.

Zudem sind mit dem Begriff Evaluation unterschiedliche Bedeutungen verbunden, was in der Praxis jedoch nicht durchgängig differenziert wird: Im Kontext etwa von Forschungsprogrammen wird mit Evaluation oder Evaluierung die wissenschaftlich fundierte Erfassung von Wirkungen bezeichnet, die durch die Implementation von Interventionen hervorgerufen werden (Herrmann et al. 2007). Andererseits wird – gerade auch im Kontext von Lehrevaluationen – Evaluation mit Bewertung gleichgesetzt. Dies ist problematisch, da eine Bewertung immer das Ergebnis einer von Individuen oder Gruppen vorgenommenen interessenbehafteten Gewichtung von Werten ist. Subjektive Werturteile sind in diesem Fall dann nicht Basis einer aus der forscherischen Distanz heraus durchgeführten Evaluation, sondern sie stellen selbst unmittelbar die Evaluation dar (im Extremfall bemisst sich gute Lehre allein daran, ob sie den Studierenden Spaß gemacht hat; vgl. Kromrey 2008).

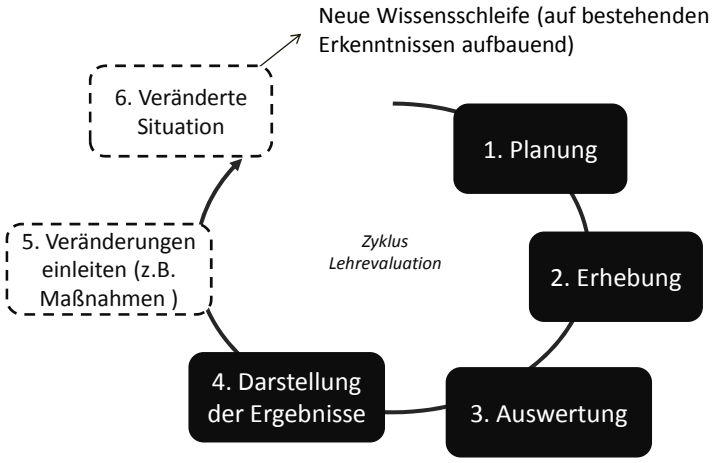
Im Gegensatz hierzu erfolgt Evaluation als Methode auf der Grundlage intersubjektiv nachvollziehbarer Verfahren und Kriterien zur Erfassung von Werten. Ziele und Bewertungskriterien werden dabei etwa durch Auftraggeber/innen, durch die Zielgruppe, beteiligte Interessengruppen, durch die Evaluator/innen selbst oder gemeinsam festgelegt (vgl. zur Evaluation als Methode und zur Evaluationsforschung Stockmann 2006; Stockmann, Meyer 2010).

Der folgende Beitrag untersucht auf der Basis einer explorativen Studie die Lehrevaluation an Hochschulen und stellt dabei den Nutzen derzeitiger Evaluationen von Lehrveranstaltungen durch standardisierte Befragungen Studierender in Frage. Im ersten Schritt konkretisieren wir die Fragestellung, aus der wir zweitens drei Hypothesen ableiten, die wir im Weiteren anhand einer Inhaltsanalyse von Fragebogenmustern studentischer Lehrkritik empirisch prüfen.

2. Zur Fragestellung: Wissensschleifen in der Lehrevaluation

Den Rahmen der Fragestellung setzt ein Projektvorhaben, das untersucht, an welcher Stelle und aus welchen Gründen so genannte Wissensschleifen in der Lehrevaluation gegebenenfalls unterbrochen werden. Der Begriff der Wissensschleife richtet sich auf einen dynamischen Prozess (der Qualitätssicherung), der im Idealfall mehrere Schritte umfasst, und zwar die Planung, Durchführung und Auswertung der Evaluation, die Erstellung eines Maßnahmenkatalogs, die Umsetzung dieser Maßnahmen, schließlich mit zeitlichem Abstand die erneute Prüfung des Status quo mit Bezug auf die vorige Evaluation und Implementation (vgl. ein ähnliches Konzept von Wissensschleifen in Bezug auf Arbeitsgestaltung bei Herrmann et al. 2007).

Abb. 1: Wissensschleife der Lehrevaluation



Unsere Annahme lautet, dass diese Schleifen meist unterbrochen werden, dass also vollständige Wissensschleifen entweder nicht hinreichend seitens der Hochschule institutionalisiert sind oder in der praktischen Umsetzung scheitern – und dass der Nutzen der Evaluationen entsprechend begrenzt bleibt. Beispielsweise könnte es sein, dass aus der studentischen Lehrkritik keine konkreten Maßnahmen, z.B. Weiterbildungsangebote oder die Studienorganisation betreffend, abgeleitet werden (Schritt 5), oder dass Lehrevaluationen eines Semesters nicht systematisch auf die Ergebnisse und Maßnahmen aus früheren Semestern bezogen werden (neue Schleife). Solche Unterbrechungen können durch folgende Ursachenkomplexe bedingt sein:

1. *Geringe Rückführbarkeit von Mängeln in der Lehrqualität auf konkrete – komplexe – Ursachen*: Die Rückführung von möglichen Missständen in der Lehre auf Ursachen (Inhalte und Didaktik der Veranstaltung selbst und/oder Kontextbedingungen) ist nur schwierig vorzunehmen. Folglich sind gezielte Verbesserungsmaßnahmen sowie die Messung ihrer »Wirksamkeit« nur bedingt möglich. Zu solchen Kontextfaktoren gehören etwa organisatorische Rahmenbedingungen (z.B. Überschneidung von Veranstaltungen, heterogenes Vorwissen) oder Merkmale der Studierenden, die ihre Bewertung beeinflussen. Es müsste z.B. unterschieden werden können, ob sich Studierende deshalb nicht für die Veran-

staltung interessieren, weil sie ganz andere Erwartungen an ihr Studium hatten, weil der Dozent[1] didaktische Mängel erkennen lässt oder deshalb, weil sie die Zahl der Teilnehmenden als zu hoch empfinden.

2. *Lehrevaluation aus Legitimationsgründen*: Eine weitere potentielle Ursache könnte darin liegen, dass die Evaluation der Lehre zumindest teilweise auch aus Legitimationsgründen erfolgt, so dass das Engagement der Beteiligten in der Hochschule, die Evaluationsschritte zu institutionalisieren, nicht an allen Hochschulstandorten gleich hoch sein dürfte.
3. *Methodische Umsetzungsprobleme*: Während sich die Argumente der kausalen Zuschreibung und des Legitimationsprinzips in erster Linie auf prinzipielle Probleme der Realisierung vollständiger Wissensschleifen richten, zielt der dritte Faktor auf die praktische Umsetzung des beschriebenen Prozesses. Sind z.B. die Planung der Evaluation und die Erstellung von Erhebungsinstrumenten/Maßnahmenkatalogen auf identische Ziele und Konzepte bezogen? Reicht ein Evaluationsdesign, das auf standardisierte Befragungen setzt, aus, oder wäre ein (ergänzendes) qualitatives Design adäquater? Sind relevante Dimensionen in der Befragung berücksichtigt? (vgl. zur methodischen Kritik z.B. Döring 2006, Kromrey 1995, 2008, Moosbrugger, Hartig 2001, Rindermann 2001, Spiel 2001).

In diesem Beitrag wird nun ein Ausschnitt potentieller Unterbrechungen von Wissensschleifen betrachtet. Die zentrale Forschungsfrage lautet: Was messen die Befragungen zu Lehrveranstaltungen von Studierenden (und was messen sie nicht)? Welche Aspekte von Lehrqualität erfassen die standardisierten Befragungen? Sind sie geeignet, das Wissen über die Qualität von Hochschullehre zu erweitern? Kann die Befragung so ausgewertet werden, dass aus geäußerter Kritik systematisch Verbesserungsmöglichkeiten für die Dozenten ableitbar sind? Der Blick richtet sich damit insbesondere auf die Erhebung, insbesondere das Erhebungsinstrument (Schritt 2 in Abb. 1) mit Bezug auf die Planung (Schritt 1) und die Auswertung (Schritt 3). Konkreter Untersuchungsgegenstand der Studie sind standardisierte Fragebögen zur Evaluation von Lehrveranstaltungen. Der folgende Abschnitt stellt diese zunächst vor, bevor dann Hypothesen zur Fragestellung sowie ihre Operationalisierung entwickelt werden

— — — — — —

1 Im Folgenden wird aus Gründen der Lesbarkeit meist vom »Dozenten« gesprochen, dabei sind männliche und weibliche Lehrende eingeschlossen.

3. Untersuchungsdesign

3.1. Lehrveranstaltungsevaluation mit standardisierten Fragebögen

In diesem Beitrag werden die Inhalte standardisierter Fragebögen, die zur Erhebung der Zufriedenheit mit einer Lehrveranstaltung dienen, untersucht. Es handelt sich um Fragebögen, die eine Hochschule bzw. Fachrichtung den Dozenten zur Verfügung stellt. In der Regel werden solche Bögen zum Ende einer Lehrveranstaltung an die Studierenden ausgehändigt und anonym ausgefüllt. Oftmals unterstützen technische Programme wie Evaprof (Ernst 2008) und Evasys (Fragebögen werden eingescannt) die Datenauswertung.

Die Verbindlichkeit des Einsatzes sowie die Inhalte der Fragebögen variieren je nach Hochschule. Bestandteil ist in der Regel ein Block, der die Zufriedenheit mit der Veranstaltung anhand verschiedender Aspekte abfragt, z.B. auf einer Fünferskala von »stimme voll zu« bis »stimme gar nicht zu«. Exemplarisch sind dies etwa Aspekte zum Umfang und zur Struktur des Stoffes, zur Art und zum Tempo der Vermittlung (z.B. »die Inhalte werden in einem für mich angemessenen Tempo vermittelt«) , zu eingesetzten Hilfsmitteln (z.B. Power Point) und Arbeitsmaterialien (z.B. »die Arbeitsmaterialien waren hilfreich«), zur Fairness und Erreichbarkeit des Dozenten (z.B. »der Dozent geht auf die Beiträge und Fragen ein«, »der Dozent schafft eine angenehme Atmosphäre«), zur Größe des Seminarraums und zur Erweiterung der eigenen Kompetenzen durch die Veranstaltung (weitere Beispiele in Ernst 2008, S. 76-79). Weitere Frageblöcke sind nicht Bestandteil jedes Fragebogens, sondern variieren in Umfang und Art, beispielsweise zur Gesamteinschätzung der Veranstaltung (z.B. »es hat sich für mich gelohnt, die Veranstaltung zu besuchen«) und des Lernerfolgs oder mit Angaben zur Person (Geschlecht, Semesteranzahl etc.) und Rahmenbedingungen (z.B. Zeit für Vor- und Nachbereitung) oder auch als offene Frage zu Lob und Kritik.

Legt man ein allgemeines Input-Output-Modell zugrunde, misst die studentische Lehrkritik nur einen spezifischen, gleichwohl wichtigen Ausschnitt von Bewertungsaspekten einer Lehrveranstaltung. Inputs werden prinzipiell durch äußere Bedingungen, durch die Dozenten und durch die Studierenden gesetzt. Die Evaluationsbögen erfassen davon direkt nur die Beiträge der Studierenden (z.B. Zeit für Vor- und Nachbereitung), allenfalls indirekt auch einige Beiträge des Dozenten, z.B. ob und welche Materialien er laut Auskunft der Studierenden benutzt. Auf der »objektiven«

Outputseite hängt die Berücksichtigung kurzfristiger Effekte vom Befragungszeitpunkt ab. Ein relativ später Zeitpunkt könnte erfassen, ob Prüfungsleistungen erfolgreich, ggf. mit welcher Note, erbracht wurden. Langfristige Effekte (z.B. die Nutzung erworbener Kenntnisse im weiteren Studium oder im Beruf) kann die Befragung, die sich auf eine spezifische Veranstaltung bezieht, nicht erfassen (dieser Aspekt ist eher für Absolventenbefragungen typisch). Insbesondere erschließt eine solche Befragung den »subjektiven« Output, also die Zufriedenheit bzw. Unzufriedenheit der Studierenden mit der Veranstaltung, wie sie in den oben genannten Dimensionen exemplarisch zum Ausdruck kommt.

3.2 Hypothesen

Wir gehen davon aus, dass die berücksichtigten Beurteilungsaspekte von Lehrveranstaltungen zumindest bislang oft nicht das Ergebnis einer Operationalisierung konzeptioneller Modelle sind (vgl. Döring 2006), die ggf. sogar über Hochschulen hinweg erarbeitet wurden, sondern eher pragmatischen Faktoren und Ad-hoc-Ideen folgen. Zweitens besteht wenig Konsens über die zentralen Bewertungsaspekte. Daraus leiten wir die These ab, dass diese über die Hochschulen hinweg vergleichsweise heterogen sind.

Hypothese 1 lautet somit: *Die für die Bewertung von Lehrveranstaltungen als relevant erachteten Merkmale fallen im Vergleich der Hochschulen deutlich heterogen aus.*

Es wird erwartet, dass unterschiedliche Bewertungsaspekte erfragt werden, die die Gesamtzufriedenheit konstituieren bzw. diese erklären könnten. Jedoch fehlen insbesondere alternative Erklärungen für die Bewertungen, die etwa in individuellen Merkmalen der Studierenden (z.B. ihren Erwartungen) sowie Kontextfaktoren innerhalb und außerhalb des Studiums (z.B. schlechte zeitliche Passung von Lehrveranstaltung und Nebenerwerbstätigkeit der Studierenden) zu finden wären.

Der Blick einer Evaluation, die auf die Lehrveranstaltung abzielt, würde demzufolge den Lehrveranstaltungshorizont nur wenig überschreiten, da es dazu ausgearbeiteter Annahmen über Zusammenhänge zwischen diesen Faktoren bedürfte. Wie oben angesprochen, hat die methodisch orientierte Literatur bereits seit längerer Zeit auf den Einfluss von »Störvariablen« aufmerksam gemacht, z.B. das Alter der Studierenden oder das anfängliche Interesse (z.B. Cashin 1988, Kromrey 1995, Moosbrugger/Hartig 2001,

Rindermann 2001, Spiel 2001). Döring etwa beschreibt den Verpflichtungsgrad der Veranstaltung, die Themenbeliebtheit und die aktive Beteiligung der Studierenden (z.B. Gruppenarbeiten) als drei Störmdimensionen der Lehrevaluation (2006: 13). Inwieweit es sich jeweils um einen verzerrenden Bias oder um Faktoren in einem notwendig komplexen Erklärungszusammenhang handelt, sei hier dahingestellt. Jedenfalls arbeiten Untersuchungen signifikante Einflussfaktoren heraus, die nochmals verdeutlichen, dass ihre Nichtberücksichtigung zu Fehleinschätzungen in der Evaluation führen kann (als neueres Beispiel etwa Rosar, Klein 2009, die zeigen, dass auch die physische Attraktivität des Dozenten geschlechtsspezifische Vor- bzw. Nachteile für die Bewertung der Veranstaltung mit sich bringt).

Eine empirische Bestätigung der These würde eine gravierende Kritik an der Evaluation von Lehrveranstaltungen bedeuten, weil die Gültigkeit gefundener Zusammenhänge z.B. zwischen einzelnen Bewertungsfaktoren und Gesamturteil als sehr beschränkt eingeschätzt werden müsste, wenn relevante Kontextvariablen fehlen.

Hypothese 2 lautet: *Individuelle Merkmale Studierender wie z.B. ihre Erwartungen an die Veranstaltung sowie Kontexteffekte innerhalb und außerhalb des Studiums werden als potentielle Erklärungsfaktoren für Bewertungen der Lehrveranstaltungen wenig berücksichtigt.*

Beziehen sich die Hypothesen 1 und 2 darauf, welche Inhalte in Evaluationen von Lehrveranstaltungen abgefragt werden – und welche gerade nicht, geht Hypothese 3 darauf ein, ob die Lehrpersonen Verbesserungsanregungen aus der ausgewerteten Befragung entnehmen können. Der Anspruch an die Evaluation wäre in diesem Fall kein übergreifender Vergleich der Qualität von Veranstaltungen über verschiedene Veranstaltungsformen, Dozenten, Fakultäten und Hochschulen hinweg, sondern die Herausarbeitung von Erforderlichkeiten und Möglichkeiten, die Qualität der Lehre in konkret dieser Veranstaltung bzw. konkret dieses Dozenten zu optimieren. Wir erwarten hier, dass die Einschätzung von Verbesserungsmöglichkeiten eher verhalten ausfällt, denn es ist anzunehmen, dass nur teilweise eine direkte Verbindung von geäußerten Bewertungen/Unzufriedenheiten zur Person oder zum Verhalten des Dozenten besteht und sich selbst aus solch einer Verbindung für den Dozenten nicht in jedem Fall eine konkrete Verbesserungsmaßnahme ableiten lässt. Beispielsweise muss die Aussage, man habe in dem Seminar nicht viel gelernt, nichts mit dem Dozenten und seiner Didaktik zu tun haben. Eine Formulierung wie »der Dozent hilft

mir, Antworten auf Fragen zu finden« richtet sich zwar sprachlich direkt auf die Person des Dozenten, doch lassen sich keine direkten Schlussfolgerungen für die Lehre daraus ableiten. Anders wäre eine häufige Kritik an der Erreichbarkeit des Dozenten einzuschätzen, dieses Item wäre ein Beispiel für Verbesserungsmöglichkeiten durch die Evaluation von Lehrveranstaltungen.

Hypothese 3 lautet: *Aus den Bewertungen zu Lerninhalten, vermittelten Kompetenzen, Didaktik und Merkmalen der Dozentinnen und Dozenten lassen sich nur in geringem Maße Verbesserungsmaßnahmen für die Lehrenden ableiten.*

Im Folgenden werden diese Hypothesen operationalisiert.

3.3. Operationalisierung

Die empirische Umsetzung erfolgt durch eine standardisierte Inhaltsanalyse von Fragebogenmustern, die sich auf die studentische Bewertung einzelner Lehrveranstaltungen richten. Die Items der Fragebogen werden dabei Kategorien zugeordnet, die sich aus der Operationalisierung der hier formulierten Hypothesen ergeben. Zur Optimierung der Reliabilität der Zuordnung wurden die Kategorien mit Beispielen versehen bzw. häufiger vorkommende Items wurden in einer Liste gesammelt, die sicherstellte, dass jedes dieser Items immer eine identische Codierung erhielt. Zudem wurden Codierregeln aufgestellt, um anfänglich als Zweifelsfälle bewertete Items begründet zuordnen zu können[2].

In die explorative empirische Analyse gingen 17 Fragebogenmuster ein. Es handelt sich hier nicht um eine repräsentative Auswahl von Fragebögen aller Hochschulen Deutschlands. Selbst wenn man davon ausgehen könnte, dass an allen Hochschulen normierte Muster über Semester hinweg eingesetzt würden, werden diese Bögen teilweise nicht öffentlich (d.h. auf der Homepage der Hochschule ohne Verwendung eines Zugangscodes) zur Verfügung gestellt. Für die hier vorgestellte explorative Vorstudie inner-

2 Es ist offensichtlich, dass Befunde immer auch von Kategorisierungen abhängen. In Einzelfällen wären Zuordnungen auch anders möglich gewesen – die Codierregeln ermöglichten hier allerdings konsistente und damit vergleichbare Zuordnungen. Ein Beispiel: Bei Passivkonstruktionen der Items wurde der Dozent als verstecktes Subjekt des Satzes erkannt und das Item entsprechend A7 (Vermittlung) statt z.B. A6b (Kompetenzen) zugeordnet, z.B. beim Item »Eine selbständige Auseinandersetzung mit den Lerninhalten wurde in der Veranstaltung gefördert«.

halb eines größeren Projektvorhabens haben wir uns aus pragmatischen Gründen darauf beschränkt, 16 verfügbare Bögen aus Universitäten mit mehr als 10.000 Studierenden und einen Bogen einer größeren Fachhochschule einzubeziehen.[3]

Angesichts dieser Fallauswahl ist kein Anspruch auf eine Repräsentativität der Befunde für (größere) Hochschulen in Deutschland zu erheben. Gleichwohl deckt die Analyse eine Bandbreite unterschiedlich ausgerichteter Hochschulen verschiedener Bundesländer ab und liefert damit relevante explorative Hinweise im Hinblick auf die Überprüfung der Hypothesen.

Zur Operationalisierung der Hypothesen im Einzelnen:

Hypothese 1: Heterogene Inhalte in den Fragebögen: Die zur Bewertung von Lehrveranstaltungen als relevant erachteten Merkmale werden in 18 Dimensionen unterteilt (s. Tabelle 1). Teilweise wurden diese ex ante aus potentiellen Themen zusammengestellt, die Bewertungs- und Kontextmerkmale sein können, unter anderem in Anlehnung an mögliche Inputs und (insbesondere) Outputs der Veranstaltung. Teilweise wurden die Differenzierungen der Dimensionen induktiv aus dem empirischen Material gewonnen und systematisiert. Die Fragebogenfragen wurden im nächsten Schritt diesen Kategorien zugeordnet. Dabei wurde die Trennschärfe der Kategorien z.B. durch die konzeptionelle Entscheidung erhöht, dass die Codierung auch die sprachliche Formulierung des Items im Fragebogen berücksichtigt.

Beispielsweise wurden folgende Items hinsichtlich der Einschätzung, wie interessant die Veranstaltung war, unterschiedlich zugeordnet: Die Aussage »Die Inhalte haben mich interessiert« wurden als Bewertung der Lerninhalte codiert, während das Item »Der Dozent förderte mein Interesse« der Vermittlung/Didaktik durch den Dozenten (der das Interesse gegenüber einem Zeitpunkt vor Beginn der Veranstaltung ggf. erhöht hat) zugeschlagen wurde.

— — — — — —

3 Es handelt sich um folgende Hochschulen: die Universitäten in Aachen, Bielefeld, Bochum, Darmstadt, Duisburg-Essen, Frankfurt a.M., Hannover, Heidelberg, Jena, Köln, Leipzig, Marburg, München, Münster, Trier und Wuppertal sowie die Fachhochschule für Technik und Wirtschaft Berlin. Teilweise handelt es sich nicht um Fragebogenmuster für die gesamte Hochschule, sondern für eine bestimmte Fakultät (z.B. die mathematisch-naturwissenschaftliche Fakultät in Köln), teilweise auch für bestimmte Veranstaltungsformen, z.B. eine Vorlesung.

Zur Bestimmung der Homogenität oder Heterogenität wird zunächst die Häufigkeitsverteilung aller Items aller Fragebögen auf die Kategorien deskriptiv auf auffällige Häufungen überprüft sowie innerhalb ggf. stark besetzter Kategorien auf Binnendifferenzierungen. Im zweiten Schritt werden nennenswert besetzte Kategorien auf ihre Streuung über die Fragebögen hinweg untersucht.

Tabelle 1: Inhaltliche Dimensionen der Fragebögen (Kategorie A)

Kategorie A	**Inhaltliche Dimensionen**
	Allgemeine Aspekte
A1	Merkmale Studierende, z.B. Studiengang, Geschlecht
A2	Studierverhalten, z.B. Vor-/Nachbereitung, Anwesenheit
A2a	Motiv Besuch, z.B. Gründe, Pflichtveranstaltung, Interesse
A3	Kontext Studium, z.B. passt in Stundenplan
A4	Kontext außerhalb des Studiums, z.B. Zeit f. Nebenjob
A5	Erwartungen vorab, z.B. Vorbereitung für Berufspraxis
	Bewertungsaspekte
A6a	Lerninhalte, z.B. niveauvoll, aktuell
A6b	vermittelte Kompetenzen, z.B. wiss. Arbeiten
A6c	Prüfungsaspekte (explizit), z.B. gut vorbereitet
A7	Vermittlungsaspekte/Didaktik
A8	Verhalten Dozent, z.B. fair, erreichbar
A9	Äußere Bedingungen, z.B. Raum
A10	Sonstige Zufriedenheitsaspekte
A11	verwendete Materialien (unabh. v. Bewertung)
A12	Zusammengefasste Bewertungen
A13	Gewichtung von Bewertungen
A14	Platz für eigenen Text
A15	Bewertung des Fragebogens

Hypothese 2: Merkmale Studierender/Kontexteffekte werden wenig berücksichtigt: Zur Überprüfung dieser Hypothese bedarf es gegenüber der Operationalisierung zu Hypothese 1 keiner zusätzlichen Kategorien. Die Kategorien A1-5 (s. Tabelle 1) richten sich auf individuelle Merkmale der Studierenden (A1), ihr Studierverhalten (A2), die Gründe für den Besuch (A2a), Erwartungen an die Veranstaltung (A5), Kontextaspekte innerhalb des Studiums (A3) sowie Kontextaspekte außerhalb des Studiums (A4). Es wird untersucht,

welcher Anteil an Fragen aus dem Fragebogen den Kategorien A1-A5 zugeordnet wird im Vergleich zur Anzahl aller Fragen im Fragebogen. Je kleiner der Anteil der Fragen ist, die diesen Kategorien zugeordnet werden, desto eher wird Hypothese 2 bestätigt, d.h. desto weniger werden Erklärungsfaktoren wie Kontexteffekte für die Bewertung der Lehrveranstaltung berücksichtigt. Es ist offensichtlich, dass der quantitative Anteil nur ein Indikator für die Berücksichtigung alternativer Erklärungsfaktoren für die Bewertung eine Lehrveranstaltung sein kann. In Kombination mit einer großen Heterogenität von Bewertungsfaktoren oder auch der Konzentration auf nur wenige Faktoren würde die Quantität hier jedoch durchaus einen hohen Grad an Aussagekraft erreichen, da man dann nicht davon ausgehen kann, dass einige wenige bzw. sporadisch einbezogene Erklärungsfaktoren zur Kontrolle des Zusammenhangs zwischen einzelnen Bewertungsaspekten und der Gesamtzufriedenheit hinreichend sind.

Hypothese 3: Nur wenige Verbesserungshinweise für Lehrende: Für die Operationalisierung der Hypothese 3 sind die vorgefundenen Fragebogenitems weiteren Kategorien zuzuordnen (s. Tabellen 2 und 3).

Dies ist zum einen die Kategorie »Universalität der Zufriedenheitsfragen« (B). Nur ein Teil der Fragebogenfragen ist tatsächlich universal einsetzbar (B1) oder passt deshalb zu der Veranstaltung, weil sie speziell für diese Veranstaltung oder zumindest diesen Veranstaltungstyp (z.B. eine Vorlesung) ausgewählt wurde (B3). Ein anderer Teil trifft nur auf manche Veranstaltungsformen zu oder basiert auf Unterstellungen, die nicht zwingend zutreffen müssen (B2). Beispiele sind die Items »Der Dozent geht auf Diskussionen ein« oder »der Dozent erklärt Schwieriges gut«. Diskussionen sind etwa in Vorlesungen oft eher untypisch – somit lässt sich aus Kritik an fehlenden Diskussionen dort nur wenig schlussfolgern, und ob ein Student Lerninhalte als schwierig empfindet, folgt offensichtlich keinem universalen Maßstab.

Tabelle 2: Kategorie zur Universalität (B)

Kategorie B	**Universalität der Items zur Zufriedenheit (A6-13)**
B1	universell einsetzbar
B2	nur für manche Veranstaltungen zutreffend
B3	speziell für diese (Art von) Veranstaltung eingefügt

Nur die auf alle oder die spezifische Veranstaltung zutreffenden Items (B1/B3) eignen sich somit, Verbesserungspotential abzubilden. Zu dessen Erfassung ist jedoch noch eine weitere kategoriale Unterscheidung zu treffen. Diese Unterscheidung betrifft die Bewertungsaspekte Lerninhalte, Kompetenzvermittlung, explizite Prüfungsaspekte (z.B. »ich fühle mich gut auf die Modulprüfung vorbereitet«), Vermittlungsaspekte/Didaktik, weitere Verhaltensaspekte des Dozenten und ggf. sonstige Zufriedenheitsaspekte (A6, A7, A8 und A10). Andere Zufriedenheitsaspekte sind dem Dozenten explizit (z.B. A9: äußere Bedingungen) oder implizit (z.B. A12: Gesamtbeurteilung) nicht direkt zuzuordnen.

Für die genannten Kategorien hingegen wurde nun das Verbesserungspotential überprüft, das entweder deutlich wird (C1; dies bedeutet allerdings nicht zwingend, dass es sich hier um die zentralen Zufriedenheitsaspekte handelt) oder das in stärkerem Maße diffus bleibt. Entweder ist unklar, ob die Kritik auf den Dozenten rückführbar ist (C2a), oder sie ist zwar auf seine Person zurückzuführen, aber konkrete Schlussfolgerungen für die Qualität der Lehre bleiben eher offen (C2b; s. die Beispiele in Abschnitt 3.2).

Tabelle 3: Kategorie ›Verbesserungspotential‹ (C)

Kategorie C	**Verbesserungspotential für Lehrende (für A6-8, ggf. A10 und zugleich B1/3)**
C1	deutlich erkennbar
C2a	mittelbar/nicht: unklar, ob Mangel an Dozenten liegt
C2b	mittelbar/nicht: Mangel auf Dozenten rückführbar, aber Schlussfolgerung für Verbesserungen unklar
C2c	mittelbar/nicht: Sonstiges

Für die Überprüfung der Hypothese 3 lautet die Korrespondenzregel nun: Je höher der Anteil der nicht universal einsetzbaren Zufriedenheitsfragen ist (B2) und je höher der Anteil der Items ist, die zwar universal einsetzbar sind (B1/B3), aber kein direktes Verbesserungspotential aufweisen (C2), desto eher kann man von einer Bestätigung der Hypothese und damit von einer skeptischen Beurteilung des Nutzens der hier untersuchten Evaluationsinstrumente ausgehen.

4. Ergebnisse

Hypothese 1: Heterogene Inhalte in den Fragebögen

Die 17 Bögen enthielten zwischen 11 und 46 Items (im Durchschnitt 30 Items pro Fragebogen). Hier ist also bereits eine gewisse Heterogenität auszumachen. Vergleicht man die thematischen Kategorien, zeigen sich jedoch Auffälligkeiten, die klar gegen eine gleichmäßige Verteilung der Items auf alle Kategorien sprechen.

220 der insgesamt 508 Items über alle Fragebögen hinweg (43% mit einer Spannweite von 25% bis 66%) entfallen auf Vermittlungsaspekte/ Didaktik (A7), vgl. Tabelle 4. Items dazu machen in jedem Fragebogen die am häufigsten besetzte Kategorie aus, und zugleich ist die Kategorie die einzige, die in allen 17 Fragebögen besetzt wird.

Tabelle 4: Häufigkeit der Items nach inhaltlichen Dimensionen (A)

Kategorie		*Summe Items*	*Anteil Items in %*
A	**Inhaltliche Dimensionen**	**508**	**100,0**
	Allgemeine Aspekte		
A1	Merkmale Studierende, z.B. Geschlecht	37	7,3
A2	Studierverhalten, z.B. Vor-/Nachbereitung	32	6,3
A2a	Motiv Besuch, z.B. Pflichtveranstaltung	9	1,8
A3	Kontext Studium, z.B. passt in Stundenplan	9	1,8
A4	Kontext außer Studium, z.B. Zeit f. Nebenjob	3	0,6
A5	Erwartungen vorab, z.B. Vorbereitung für Beruf	0	0
	Bewertungsaspekte		
A6a	Lerninhalte, z.B. niveauvoll, aktuell	25	4,9
A6b	vermittelte Kompetenzen, z.B. wiss. Arbeiten	25	4,9
A6c	Prüfungsaspekte (explizit), z.B. gut vorbereitet	3	0,6
A7	Vermittlungsaspekte/Didaktik	220	43,3
A8	Verhalten Dozent, z.B. fair, erreichbar	44	8,7
A9	Äußere Bedingungen, z.B. Raum	24	4,7
A10	Sonstige Zufriedenheitsaspekte	13	2,6
A11	verwendete Materialien (unabh. v. Bewertung)	2	0,4
A12	Zusammengefasste Bewertungen	32	6,3
A13	Gewichtung von Bewertungen	0	0
A14	Platz für eigenen Text	25	4,9
A15	Bewertung des Fragebogens	5	1,0

In zwei Kategorien gibt es dagegen keine einzige Zuordnung: Zum einen wird die durchaus wichtige Ebene der Erwartungen Studierender nicht erfasst (A5, s. Hypothese 2), zum anderen wird unter den Bewertungsaspekten keine Gewichtung nach deren Relevanz abgefragt. Dies könnte also allenfalls indirekt über einen Vergleich einzelner Items mit einem Gesamturteil, sofern dies zur Verfügung steht, erfolgen oder aus theoretischen Annahmen abgeleitet werden.

Betrachtet man insgesamt die relative Verteilung der 508 Items auf die 18 inhaltlichen Dimensionen, zeigt sich, dass bis auf die Vermittlungsaspekte (A7) keine Kategorie mit einem auch nur zweistelligen Anteil besetzt ist. 13 der 18 Kategorien haben sogar nur einen Anteil von unter 5%. Äußere Bedingungen (A9) etwa, die zwar nicht dem Dozenten, aber ggf. der Hochschule Aufschluss geben könnten, machen nur 4,7% der Items über alle Fragebögen hinweg aus; die Bewertung von Lerninhalten (A6a) und erlangten Kompetenzen (A6b) bleiben ebenfalls knapp unter 5%. Einen Anteil zwischen 5% und 10% haben die Kategorien Merkmale der Studierenden, Studierverhalten, Verhalten des Dozenten (außer Didaktik) und zusammengefasste Zufriedenheit.

In diesem allgemeinen Sinne sind die abgefragten Themen in Lehrevaluationen also nicht sehr heterogen, es ist ein klarer Schwerpunkt auf Vermittlungsaspekte zu konstatieren. Zwei Ergänzungen differenzieren den Befund jedoch:

a) Wie sieht die Verteilung innerhalb der am häufigsten vorkommenden Kategorie »Didaktik« (A7) aus? Es entsteht der Eindruck, dass es hier einige »mainstreams« gibt, die immer wieder vorkommen, oder auch einen Pool von möglichen Fragen, aus dem mal das eine, mal das andere berücksichtigt wird, schließlich einige, die nur ein- oder zweimal vorkommen (vgl. Tabelle 5. Konkret werden Fragen, die sich – mit unterschiedlichen Formulierungen und Schwerpunktsetzungen – auf verwendete Medien, Materialien etc. und auf die Struktur und die Ziele der Veranstaltung richten, am häufigsten gestellt (15,9% bzw. 14,5% aller Items zu A7). Das Bild aller übrigen Fragen stellt sich heterogener dar. Beispielsweise gibt es in fünf Bögen Items, die sich auf Zusammenfassungen der Lehrperson beziehen; neunmal wird die Sprache des Dozenten (z.B. spricht deutlich, akustisch verständlich) thematisiert.

Tabelle 5: Unterkategorien von A7 (Bewertung von Vermittlungsaspekten)

	absolut	*%*
klarer Aufbau, Struktur, roter Faden, klare Ziele, Anforderungen	32	14,5
Medien, Tafelanschrieb, Materialien, Literaturangaben ausreichend/hilfreich	35	15,9
Dozent erklärt Schwieriges, kann gut vermitteln, verständlich, leicht zu folgen	14	6,4
Dozent vermittelt anschaulich, verwendet Beispiele	9	4,1
Dozent geht auf Fragen, Anregungen, Diskussionen, Wünsche ein	15	6,8
Tempo	10	4,5
Sprache, z.B. spricht deutlich	9	4,1
Dozent wirkte gut vorbereitet	6	2,7
Bezüge zu Aktuellem/zur Praxis (5/10)	15	6,8
Dozent fördert Beteiligung, Mitdenken, selbst. Arbeiten, kritische Auseinandersetzung, Interesse	18	8,2
Dozent gibt Zusammenfassungen	5	2,3
Sonstiges	52	23,6
SUMME (über alle 17 Fragebögen)	**220**	**100,0**

b) Wenn die zwölf Kategorien untersucht werden, deren Anteil zwischen 1% und 10% beträgt, fällt deren Streuung im Vergleich der 17 Fraugebögen zwar vergleichsweise gering aus. Das ist jedoch auch bedingt durch den insgesamt geringen Anteil der Items an allen 508 Items. Der Abstand zwischen dem ersten und dem dritten Quartil beträgt in sieben Fällen unter fünf Prozentpunkten, in fünf Fällen aber immerhin zwischen 8,7 und 10,4 Prozentpunkten (die Orientierung am Quartilsabstand wurde in diesem kleinen Sample gewählt, um »Ausreißern« kein unangemessen hohes Gewicht zu geben). Heterogenität zeigt sich also zwar nicht in absolut großen Streuungen, aber im Einzelnen in gewissen Beliebigkeiten, die zumindest zum Ausdruck bringen, dass die Operationalisierung vermutlich keinem übereinstimmenden Konzept von Lehrevaluation folgte. So werden etwa Eigenschaften des Dozenten in zwei Bögen gar nicht erfasst, in den anderen gibt es Fragen dazu bis zu einem Maximum von sieben Fragen, die hier zugeordnet wurden. Ähnliches lässt sich etwa für Lehrinhalte (zwischen 0 und 14% der Items pro Fragebogen) und Kompetenzen (zwischen 0 und 16%) sagen. Welcher konzeptionelle Stellenwert diesen Dimensionen zukommt, wird also von den für die Lehrevaluation Verantwortlichen nicht annähernd übereinstimmend eingeschätzt.

Es lässt sich zusammenfassen, dass Vermittlungsaspekte den Schwerpunkt der studentischen Lehrveranstaltungsbewertung ausmachen. Insofern bestätigt sich Hypothese 1, die eine große Heterogenität an Kategorien postuliert, nicht. Betrachtet man die in mittlerem Ausmaß vorkommenden Kategorien sowie Unterkategorien der Kategorie »Vermittlungsaspekte«, lässt sich allerdings durchaus von einer gewissen Beliebigkeit sprechen, mit der Bewertungsaspekte als zentral oder marginal betrachtet werden. Somit ist auch eine Vergleichbarkeit der studentischen Lehrkritik über verschiedene Fakultäten oder Hochschulen hinweg nicht möglich, selbst wenn Kontextfaktoren Berücksichtigung fänden, was im Folgenden anhand von Hypothese 2 überprüft wird.

Hypothese 2: Merkmale Studierender/Kontexteffekte werden wenig berücksichtigt
Zur Überprüfung der Hypothese 2 ist zu klären, wie hoch der Anteil der Items ist, die externe Erklärungsfaktoren für den Grad der Zufriedenheit mit der Veranstaltung erfassen, etwa Studierendenmerkmale oder Kontexteffekte (A1-A5, s. Tabelle 6).

Tabelle 6: Häufigkeit der allgemeinen inhaltlichen Kategorien (A1-A5)

Kategorie		*Vorkommen in … Bögen*	*Summe Items*	*Anteil Items in %*
A1	Merkmale Studierende	14	37	7,3
A2	Studierverhalten	13	32	6,3
A2a	Motiv Besuch	9	9	1,8
A3	Kontext Studium	5	9	1,8
A4	Kontext außer Studium	2	3	0,6
A5	Erwartungen	0	0	0,0
Gesamt		**17 Bögen**	**90/508**	**17,7**

Insgesamt ist der Anteil der Erklärungsalternativen für Zufriedenheit nicht sehr hoch; dies bezieht sich insbesondere auf die Kategorien A3 bis A5. Das heißt: Erwartungen werden im herangezogenen Sample gar nicht, Kontextaspekte nur sehr selten erhoben (in nur 9 Items, verteilt auf 5 der 17 Fragebögen, geht es um den Kontext innerhalb des Studiums, in 3 Items in 2 Fragebögen um solche außerhalb des Studiums).

Das Ergebnis lautet somit, dass wichtige potentielle Erklärungsfaktoren für die Zufriedenheit mit einer Lehrveranstaltung in den meisten Fällen kaum abgefragt werden. Man könnte sich etwa einen Studierenden vorstellen, der

am Dozenten, seiner Didaktik und den äußeren Bedingungen wenig auszusetzen hat, der jedoch z.B. die Theorien des Faches schlicht langweilig, nicht unterhaltsam findet und auch nicht einsieht, wofür er die Theorien später im Beruf gebrauchen könnte. Selbst wenn diese Bewertung im Beurteilungsteil unter Kategorie 6a (Lerninhalte) abgebildet würde, bliebe dahingestellt, ob die Themenauswahl innerhalb des großen Themas »Theorien« Abhilfe schaffen könnte oder ob der Studierende schon vorab besser darüber informiert werden sollte, was ihn in einem Hochschulstudium erwartet – und dieser evaluatorische Mangel ergibt sich daraus, dass seine Erwartungen nicht in die Analyse eingingen.

Am ehesten werden die Merkmale von Studierenden und das Studierverhalten im Evaluationsbogen thematisiert; diese kommen mit 7,3 bzw. 6,3% der Items noch vergleichsweise häufig vor. Es sei aber dahingestellt, ob eine weitergehende Erklärungskraft aus einem Befund ableitbar wäre, der z.B. die Aussage treffen könnte, dass Frauen zufriedener sind als Männer in einer Veranstaltung, an der deutlich mehr Männer bzw. mehr Frauen teilnehmen. Es bedarf demnach eines Operationalisierungskonzepts, das Studierendenmerkmale bzw. Studierverhalten systematisch mit Bewertungsaspekten in einen Zusammenhang stellt. Ob es dieses Konzept gibt, kann eine Analyse der Fragebögen nicht abschließend beantworten, hier sind andere Erhebungsinstrumente (z.B. eine Befragung der an Hochschulen für Lehrevaluation Verantwortlichen) anzuwenden.

Hypothese 3: Nur wenige Verbesserungshinweise für Lehrende

Zur Hypothese 3 ist zu untersuchen, welcher Anteil der Zufriedenheitsfragen (ab Dimension A6 in Tabelle 4) universale Anwendung finden kann und ob unter diesen, sofern sie sich auf den Dozenten und seine Didaktik, auf Inhalte und auf vermittelte Kompetenzen beziehen (d.h. auf die Kategorien A6-A8, A10), ein direktes Verbesserungspotential für den Dozenten abgelesen werden kann.

Tabelle 7: Häufigkeit in der Kategorie Universalität (B)

Kategorie		*Summe Items*	*Anteil Items in %*
B	**Universalität der Items zur Zufriedenheit (A6-13)**	**382**	**100,0**
B1	universell einsetzbar	279	73,0
B2	nur für manche Veranstaltungen zutreffend	93	24,3
B3	speziell für diese (Art von) Veranstaltung eingefügt	10	2,6

Immerhin knapp ein Viertel der Zufriedenheitsfragen (93 von 382, 24,3%) ist nicht universal anzuwenden. In einigen Fällen ist eine Antwortmöglichkeit wie etwa »keine Angabe« vorgesehen, doch wird offenbar häufig übersehen, dass eine eindeutige Zuordnung von angekreuzter Antwort und einer Beurteilung nicht möglich ist. Wenn beispielsweise abgefragt wird, ob die Veranstaltung tätigkeitsrelevantes Wissen vermittelt habe, kann eine Verneinung entweder darauf hindeuten, dass dies einen Nachteil der Veranstaltung darstellt oder andererseits, dass die Veranstaltung nicht auf unmittelbare Praxisverwertung angelegt war (z.B. »Philosophische Theorien in Geschichte und Gegenwart«). Ein zweites Beispiel: »Didaktische Hilfsmittel (z.B. Flipchart) wurden sinnvoll eingesetzt«. Dieses Item ist mehrdeutig, es richtet sich zum einen auf den Einsatz von Hilfsmitteln und zum anderen auf ihre Bewertung. Die Antwort »selten« oder »nie« kann sich auf beide Deutungen beziehen. Wenn fast ein Viertel der Zufriedenheitsfragen mit diesen Uneindeutigkeiten behaftet ist, schränkt dies den Ertrag der Befragungen für die Bewertung von Lehrveranstaltungen deutlich ein.

Tabelle 8: Häufigkeit in der Kategorie Verbesserungspotenzial (C)

Kategorie		*Summe Items*	*Anteil Items in %*
C	**Verbesserungspotential für Lehrende (für A6-8, ggf. A10 und zugleich B1/3)**	**234**	**100,0**
C1	Potential deutlich erkennbar	110	47,0
C2 (Summe)	Potential nicht deutlich erkennbar	124	53,0
C2a	mittelbar/nicht: unklar, ob Mangel an Dozenten liegt	74	31,6
C2b	mittelbar/nicht: Mangel auf Dozenten rückführbar, aber Schlussfolgerung für Verbesserungen unklar	48	20,5
C2c	mittelbar/nicht: sonstiges	2	0,9

Kann der Dozent zumindest von den übrigen Bewertungsfragen, die immerhin die Mehrheit darstellen, konkrete Verbesserungen für künftige Veranstaltungen ableiten? Von 234 Items, die in die Kategorie »Verbesserungspotential« (C) einzuordnen waren, wurden über alle Bögen hinweg über die Hälfte (53% mit einer Spannweite zwischen 36% und 75%) der Unterkategorie zugeordnet, die allenfalls mittelbares Verbesserungspotential für den Dozenten anzeigt. Dabei entfallen 31,6% auf die Kategorie C2a (es ist unklar, ob die Kritik an Handlungsweisen des Dozenten liegt) und 20,5% auf die Kategorie C2b (die Kritik ist der Formulierung nach auf den Dozenten zurückzuführen, aber die Schlussfolgerungen für die Lehre

bleiben diffus). Beispiele für unklare Zuordnungen zum Dozenten (C2a) sind »die Lehrveranstaltung/der Dozent förderte mein Interesse« (dies hängt unter anderem auch vom Interesse für das Thema vor dem Besuch der Veranstaltung ab) oder »die Inhalte sind relevant« (welchen Maßstab setzt der Studierende hier an?). Beispiele für unklare Folgerungen der Kritik am Dozenten (C2b) lauten »der Dozent wirkt kompetent« oder »der Dozent hat Interesse an den Studierenden«. Der Dozent erhält zwar potentiell eine Rückmeldung über wahrgenommene Mängel, doch sind Konsequenzen nur schwer zu ziehen – was genau soll er etwa tun, damit seine Kompetenz, die er selbst aus seiner fachlichen Qualifikation ableiten mag, für die Studierenden sichtbarer wird?

Dieser Indikator weist noch deutlicher darauf hin, dass erstaunlich häufig kein systematischer Bezug zwischen Bewertungsitems und Verbesserungspotential besteht. Offen bleiben muss an dieser Stelle, inwieweit solch ein Bezug für eine große Bandbreite an Evaluationsfaktoren herstellbar ist, inwieweit das Instrument also geeignet ist, um Verbesserungspotential herauszuarbeiten, das über Aspekte wie Pünktlichkeit, deutliche Aussprache und das Formulieren klarer Anforderungen etc. hinausgeht. Die explorative Diagnose des Ist-Zustands muss jedenfalls zunächst einmal skeptisch ausfallen.

5. Fazit

Der Beitrag befasste sich mit der Frage, welchen Nutzen Evaluationen der Hochschullehre erbringen können. Zu diesem Zweck untersuchten wir in einer explorativen Studie Fragebögen, die sich an Studierende zur Lehrveranstaltungsevaluation an Hochschulen richten. Insbesondere wurde untersucht, a) welche inhaltlichen Aspekte die Befragungen beinhalten und b) inwiefern sich daraus ein Verbesserungspotential für Dozentinnen und Dozenten ergibt.

Zusammengefasst hat die explorative Überprüfung folgende Befunde erbracht:

– Die Fragebögen zur studentischen Lehrkritik konzentrieren sich auf Vermittlungsaspekte. Darüber hinaus gibt es zwar Items, die in mehreren Bögen erfragt werden, doch lässt sich zumindest aus den Fragebögen nicht auf einen deutlichen Konsens über zentrale Evaluationsfakto-

ren schließen. Hypothese 1 hat sich damit teilweise bestätigt, d.h. es ist – abgesehen vom generellen Fokus auf Vermittlungsaspekte – nicht von homogenen Schwerpunkten und Themen in den Fragebögen auszugehen.

- Insbesondere Kontexteffekte innerhalb und außerhalb des Studiums und Erwartungen Studierender werden als potentielle Erklärungsfaktoren für die Bewertung von Lehrveranstaltungen kaum berücksichtigt (d.h. die Daten bestätigen Hypothese 2). Dies ist umso erstaunlicher, als nicht erst die jüngste (methodisch orientierte) Literatur zum Thema auf den Einfluss solcher »Störvariablen« aufmerksam macht. Individuelle Merkmale und das Studierverhalten werden leicht häufiger thematisiert, jedoch bleibt der Nutzen, den Auswertungen der Evaluationsbögen aus den Verteilungen dieser Items ziehen, hier (zwangsläufig) offen.
- Zu einem nennenswerten Anteil werden in Bezug auf Inhalte, Didaktik, vermittelte Kompetenzen und Merkmale des Dozenten Items formuliert, aus denen sich potentielle Verbesserungsmaßnahmen für Lehrende nicht direkt ableiten lassen (wie in Hypothese 3 postuliert).

Einschränkend ist anzuführen, dass es sich bei diesen Befunden erstens um eine explorative Studie ohne Anspruch auf Repräsentativität handelt, wenngleich eine gewisse Breite an Hochschulen einbezogen werden konnte. Zweitens bildeten die Fragebogenitems zur studentischen Lehrkritik die Instanz zur Überprüfung der Hypothesen und nicht z.B. Aussagen der an der Hochschule für Evaluation zuständigen Personen (im Rektorat, in der Verwaltung etc.) oder weitere Materialien (z.B. Evaluationskonzepte).[4] Möglicherweise ist die standardisierte Befragung Teil eines umfangreicheren Evaluationskonzepts mit weiteren Erhebungsinstrumenten (wie z.B. an der RWTH Aachen und das Berliner Evaluationsinstrument, vgl. Braun et al. 2008). Auch dann stellt sich allerdings die Anschlussfrage, wie die Ergebnisse der standardisierten Befragung Studierender mit den Ergebnissen aus anderen Datenerhebungen zueinander in Bezug gesetzt werden.

Welche Schlussfolgerungen lassen sich aus diesen Befunden für die Lehrevaluation an Hochschulen ziehen?

Angestrebte Wissensschleifen werden oft schon im Zuge der ersten Schritte unterbrochen, insofern die Datenerhebung zum einen vermutlich

— — — — — —

4 Solche Erweiterungen sind in unserem geplanten Projektvorhaben vorgesehen.

keinem Konzept folgt, das vollständige Wissensschleifen im Blick hätte und in entsprechende Institutionalisierungen eingebunden wäre. Zum anderen zieht sie nicht systematisch Auswertungen nach sich, die zu gezielten Diskussionen und Einleitungen von Verbesserungsmaßnahmen beitragen könnten. Damit führen die standardisierten Fragebögen als Instrument der Lehrevaluation tendenziell zu einem Datenfriedhof und tragen nicht zu einer Optimierung der Qualität der Lehre bei. Dieses Problem wurde bislang nicht derart reflektiert, dass es sich in aktuell verwendeten Evaluationsbögen niedergeschlagen hätte.

Die fundierte Diagnose dieser Unterbrechung von Wissensschleifen steht im Zentrum der vorliegenden Analyse von Evaluationsfragebögen. Zumindest ausblickend ist weiterhin zu klären, welchen Beitrag die Studie leisten kann, um die Ursachen dieser Unterbrechung von Wissensschleifen zur Qualitätsverbesserung der Lehre aufzuklären. Oben wurden drei mögliche Ursachenkomplexe erläutert: das Problem der kausalen Zuordnung von Missständen zu Gründen hierfür, die ggf. vorherrschende Legitimationsfunktion von Evaluationen sowie methodische Umsetzungsprobleme. Zur Prüfung der Legitimationsfunktion von Evaluationen müssten weitere Instrumente herangezogen werden, die unter anderem die subjektive Sicht der beteiligten Akteure an den Hochschulen einbeziehen. Zu den anderen beiden Ursachenkomplexen lassen sich aus den Befunden erste begründete Schlussfolgerungen ziehen: Zumindest die methodische Umsetzung von Lehrevaluationen stellt derzeit an vielen Hochschulen ein Problem dar, das dringend der Reflexion und Bearbeitung bedarf. Eine systematische Institutionalisierung vollständiger Wissensschleifen und in der Folge ein stringentes Evaluationskonzept, das erstens eine komplexe Datenerhebung unter Berücksichtigung vielfältiger Erklärungsfaktoren der Bewertung von Lehrveranstaltungen einschließt, zweitens standardisierten Befragungen eine definierte Rolle im Evaluationsprozess zuweist und das schließlich die Bewertungen Studierender nicht ohne Begründung mit der Evaluation selbst gleichsetzt, ist für die Feststellung des Status quo der Lehrqualität und ggf. ihrer Verbesserung das anzustrebende Ziel. Erst auf der Grundlage einer Analyse solch komplex angelegter Evaluationen ließe sich dann beurteilen, ob zusätzlich ein prinzipielles Problem der kausalen Zuordnung von Missständen zu Gründen hierfür vorliegt. Allein auf der Basis standardisierter Befragungen erscheint eine solche Zuordnung zurzeit jedoch kaum möglich.

Literatur

Ammon, S.; Heineke, C., Selbmann, K. (Hg.) 2007: Wissen in Bewegung: Vielfalt und Hegemonie in der Wissensgesellschaft. Weilerswist: Velbrück.

Bittlingmayer, U., Bauer, U. (Hg.) 2006: Die »Wissensgesellschaft«. Mythos, Ideologie oder Realität? Wiesbaden: VS.

Braun, E., Gusy, B., Leidner, B., Hannover, B. 2008: Kompetenzorientierte Lehrevaluation – Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). Diagnostica, Jg. 54, H. 1, 30–42.

Cashin, W. E. 1988: Student ratings of teaching: A summary of the research. IDEA Paper No. 20. Manhattan, Ks.: Kansas State University, Center for Faculty Evaluation and Development.

Döring, N. 2006: Für Evaluation und gegen Evaluitis. In B. Berendt, H.-P. Voss und J. Wildt (Hg.), Neues Handbuch Hochschullehre. Lehren und Lernen effizient gestalten, 2., überarb. Auflage, Stuttgart: Raabe, Beitrag I 1.7. (Loseblatt-Sammlung).

Ernst, St. 2008: Manual Lehrevaluation. Wiesbaden: VS.

Frey, B. 2006: Evaluitis – eine neue Krankheit. Working paper No. 293, Institute for Empirical Research in Economics. Zürich.
http://homepage.univie.ac.at/Eveline.Christof/evaluation09/Evaluitis.pdf, Download am 29.07.2010.

Herrmann, Th.; Jahnke, I., Klick, H., Skrotzki, R. 2007: Ex-Post und Ex-ante Evaluation des BMBF-Rahmenkonzeptes »Innovative Arbeitsgestaltung, Zukunft der Arbeit«: Zukunft eines Forschungsprogramms. In D. Streich, D. Wahl (Hg.), Innovationsfähigkeit in einer modernen Arbeitswelt. Frankfurt a.M.: Campus, 501–528.

Kromrey, H. 1995: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In P. Mohler (Hg.), Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung, Münster: Waxmann, 2. Aufl., 105–127.

Kromrey, H. 2008: Wissenschaftstheoretische Anforderungen an empirische Forschung und die Problematik ihrer Beachtung in der Evaluation. In K.-S. Rehberg (Hg.), Die Natur der Gesellschaft. Verhandlungen des 33. Kongresses der DGS in Kassel 2006. Frankfurt a.M.: Campus, 1923–1932.

Moosbrugger, H., Hartig, J. 2001: Zur Bedeutung von individuellen und institutionellen Studienbedingungen für die vergleichende Evaluation der Lehre. In U. Engel (Hg.), Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre. Frankfurt/New York: Campus, 49–60.

Pohlenz, Ph. 2008: Lehrevaluation und Qualitätsmanagement – neue Anforderungen für die Hochschulsteuerung. Sozialwissenschaften und Berufspraxis, Jg. 31, H. 1, 66–78.

Rindermann, H. 2001: Die studentische Beurteilung von Lehrveranstaltungen – Forschungsstand und Implikationen. In Ch. Spiel (Hg.), Evaluation universitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck. Münster: Waxmann, 61–88.

Rosar, U., Klein, M. 2009: Mein(schöner)Prof.de. Die physische Attraktivität des akademischen Lehrpersonals und ihr Einfluss auf die Ergebnisse studentischer Lehrevaluationen. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Jg. 61, H. 4, 621–645.

Spiel, Ch. 2001: Der differentielle Einfluss von Biasvariablen auf studentische Lehrveranstaltungsbewertungen. In: U. Engel (Hg.), Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre. Frankfurt/New York: Campus, 61–82.

Stockmann, R. (Hg.) 2006: Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann.

Stockmann, R., Meyer, W. 2010: Evaluation. Eine Einführung. Opladen, Bloomfield Hills: B. Budrich (UTB).

Tänzler, D., Knoblauch, H. (Hg.) 2006: Zur Kritik der Wissensgesellschaft. Konstanz: UVK.

Wehling, P. 2007: Die Politisierung des Nichtwissens: Vorbote einer reflexiven Wissensgesellschaft? In S. Ammon, C. Heineke, K. Selbmann, A. Hintz (Hg.), Wissen in Bewegung. Vielfalt und Hegemonie in der Wissensgesellschaft. Weilerswist: Velbrück, 221–240.

Würmseer, G. 2010: Hochschulen im Identitätswandel (Arbeitstitel). Dissertation. Wiesbaden: VS. In Vorbereitung.