

WESTFÄLISCHE
WILHELMS-UNIVERSITÄT
MÜNSTER



Diskussionspapier des Instituts für Organisationsökonomik

8/2012

Effekte von Erhebungsart und -zeitpunkt auf studentische Evaluationsergebnisse

Laura Lütkenhöner

Discussion Paper of the
Institute for Organisational Economics

**Diskussionspapier des
Instituts für Organisationsökonomik
8/2012**

August 2012

ISSN 2191-2475

**Effekte von Erhebungsart und -zeitpunkt
auf studentische Evaluationsergebnisse**

Laura Lütkenhöner

Zusammenfassung

Es gibt keine einheitliche Praxis hinsichtlich der methodischen Konzeption und des Zeitpunkts der Datenerhebung bei studentischen Lehrveranstaltungsevaluationen. Um den Einfluss solcher Aspekte zu untersuchen, wurden in 18 Wirtschaftsvorlesungen der Westfälischen Wilhelms-Universität Münster neben der offiziellen schriftlichen Befragung der Studierenden (Papier-Evaluation) zu mehreren Zeitpunkten Online-Befragungen durchgeführt. Signifikante systematische Unterschiede wurden vor allem zwischen Bewertungen, die zu verschiedenen Zeitpunkten erhoben wurden, gefunden. Im gleichen Zeitraum durchgeführte Online- und Papier-Evaluationen führten trotz deutlicher Unterschiede in den Rücklaufquoten zu hochgradig korrelierten Ergebnissen in fast allen Bewertungskategorien.

JEL-Codes: A20, A22, C23, I21

Effects of Survey Method and Time on Student Evaluations

Abstract

There is no unique method for student evaluations of teaching (SET) that also take place at different points of time. In order to analyse the influence these aspects have on the results of SET, 18 economic lectures at the University of Münster were evaluated not only by distributing official paper surveys but also by asking the students to fill out online surveys at several points of time. The empirical study has revealed significant systematic differences especially between SET results evaluated at different points of time. If assessed at the same point of time SET results evaluated by means of online-surveys are mostly highly correlated with SET results evaluated by means of paper surveys despite notable differences in the response rate.

Im Internet unter:

http://www.wiwi.uni-muenster.de/io/forschen/downloads/DP-IO_08_2012.pdf

Westfälische Wilhelms-Universität Münster
Institut für Organisationsökonomik
Scharnhorststraße 100
D-48151 Münster

Tel: +49-251/83-24303 (Sekretariat)
E-Mail: io@uni-muenster.de
Internet: www.wiwi.uni-muenster.de/io

Effekte von Erhebungsart und -zeitpunkt auf studentische Evaluationsergebnisse*

1. Einleitung

Die Landeshochschulgesetze verpflichten Hochschulen dazu, regelmäßig ihre Aufgabenerfüllung im Bereich der Lehre zu überprüfen, und sie sehen zum Teil explizit eine Mitwirkung von Studierenden an der Bewertung von Lehrveranstaltungen vor (vgl. u. a. Art. 10 Abs. 3 S. 1 Bayerisches Hochschulgesetz, § 3a Abs. 3 Gesetz über die Hochschulen des Landes Mecklenburg-Vorpommern, § 8 Abs. 3 Thüringer Hochschulgesetz). Studentische Evaluationen in Form von schriftlichen Befragungen von Studierenden zu ihrer individuellen Einschätzung der Qualität besuchter Lehrveranstaltungen sind dabei gängige Praxis. Die Ergebnisse werden im Hochschulbetrieb aber nicht nur, wie gesetzlich vorgesehen, zur Qualitätsentwicklung und -sicherung genutzt (vgl. u. a. § 7 Abs. 2 S. 1 Hochschulgesetz Nordrhein-Westfalen), sondern wirken zunehmend auch in andere Bereiche des Hochschulbetriebs hinein (vgl. Meinefeld 2010, S. 314). Wie Wettern (2008, S. 30) berichtet, spielen sie an der Universität Osnabrück bei Bleibeverhandlungen und für die Verlängerung von Juniorprofessuren eine Rolle und haben Einfluss auf den leistungsbezogenen Teil persönlicher W-Besoldungen. Teilweise werden Ergebnisberichte studentischer Evaluationen auch Bewerbungen um Professuren beigelegt (vgl. Meinefeld 2010, S. 314). Außerdem werden sie zur Auszeichnung von Lehrenden auf Fakultäts-, Universitäts-, Landes- und Bundesebene genutzt (vgl. Meinefeld 2010, S. 309). Diese Beispiele zeigen, dass studentische Evaluationen einen hohen Stellenwert im Hochschulalltag haben. Dieser hat im Zuge der Einführung von Bachelor- und Masterstudiengängen und „der damit verbundenen, deutlich verstärkten und länderübergreifenden Konkurrenzsituation zwischen den Hochschulen“ (Wettern 2008, S. 29) eher noch zugenommen.

Studentische Evaluationen sind aber nicht unumstritten. So zeigen zahlreiche Studien, dass die Ergebnisse von mehr oder weniger sachfremden Faktoren beeinflusst werden, zum Beispiel von Noten (vgl. u. a. Brockx/Spooren/Mortelmans 2011) oder von der Attraktivität des Lehrenden (vgl. u. a. Felton/Koper/Mitchell/Stinson 2008). Zwei andere Aspekte von mög-

* Herrn Prof. Dr. Alexander Dilger danke ich für wertvolle Ideen, Anregungen und Hinweise bei der diesem Diskussionspapier zugrunde liegenden Masterarbeit sowie Änderungen und Kommentare bei der Erstellung dieses Diskussionspapiers. Ihm und den anderen Professoren des Centrums für Interdisziplinäre Wirtschaftsforschung der Westfälischen Wilhelms-Universität Münster, namentlich Herrn Prof. Dr. Thomas Apolte, Herrn Prof. Dr. Dietmar Krafft sowie Herrn Prof. Dr. Christian Müller, danke ich außerdem für ihr Einverständnis, dass ich unter den Teilnehmern ihrer Lehrveranstaltungen Befragungen durchführen durfte. Schließlich danke ich allen Studierenden, die an den Befragungen teilgenommen haben. Für alle Inhalte und möglicherweise verbliebene Fehler bin selbstverständlich ausschließlich ich selbst verantwortlich.

licherweise erheblicher Bedeutung, die bisher aber nur vereinzelt untersucht wurden (u. a. von Pruitt/Dicks/Tilley 2011, Meinefeld 2010, Avery/Bryant/Mathios/Kang/Bell 2006, Layne/DeCristoforo/McGinty 1999 und Kohlan 1973) sind die Erhebungsart und der Evaluationszeitpunkt. Vor dem Hintergrund, dass studentische Evaluationen immer häufiger auch für Leistungsvergleiche der Lehrenden verwendet werden, erscheint es dringend geboten, den Einfluss dieser beiden Faktoren näher zu untersuchen.¹ Da Online-Evaluationen im Gegensatz zu papiergestützten Evaluationen in Abwesenheit des Dozenten erfolgen und die Teilnahme für die Studenten Aufgabe von frei verfügbarer Zeit bedeutet, wären systematische Unterschiede zwischen den Ergebnisse beider Erhebungsarten durchaus denkbar. Ein systematischer Einfluss des Evaluationszeitpunkts ist schon deshalb nicht auszuschließen, weil in universitären Lehrveranstaltungen oftmals keine Anwesenheitspflicht herrscht, so dass die Zusammensetzung der Teilnehmer über das Semester hinweg variieren kann. Auch ist vorstellbar, dass sich die Wahrnehmung der Lehrqualität von Studierenden zu verschiedenen Zeitpunkten unterscheidet (sie könnte z. B. kurz vor der Klausur schlechter ausfallen). Inwiefern derartige Tendenzen empirisch beobachtbar sind, wird im Rahmen dieser Studie anhand von Wirtschaftsvorlesungen der Westfälischen Wilhelms-Universität Münster getestet, die jeweils zu mehreren Semesterzeitpunkten evaluiert wurden. Die Ergebnisse eigener Online-Befragungen werden dabei nicht nur miteinander, sondern auch mit den Ergebnissen der offiziellen Lehrevaluationen, bei denen Fragebögen im Hörsaal verteilt wurden, verglichen.

Im nachfolgenden zweiten Abschnitt wird der Ablauf der Evaluation von wirtschaftswissenschaftlichen Lehrveranstaltungen an verschiedenen deutschen Universitäten thematisiert. Der dritte Abschnitt gibt einen Überblick über den Stand der Forschung. Im vierten Abschnitt werden die Erhebung und Aufbereitung der verwendeten Daten sowie die Vorgehensweise bei der Analyse des Einflusses von Erhebungsart und -zeitpunkt beschrieben. Dabei wird auch die Idee erläutert, Unterschiede in der Bewertung von Teilaspekten einer Vorlesung dazu zu verwenden, um die Differenziertheit von studentischen Evaluationen zu operationalisieren. Die Ergebnisse werden im fünften Abschnitt präsentiert und im sechsten Abschnitt diskutiert. Die Arbeit schließt mit einem Fazit und Anregungen für weitere Forschung.

¹ Ungeachtet etwaiger derartiger Einflüsse sind solche Vergleiche von Lehrenden grundsätzlich problematisch, da bessere Evaluationsergebnisse nicht inhaltlich besserer Lehre entsprechen müssen und sogar eine negative Korrelation denkbar ist.

2. Ablauf studentischer Evaluationen an wirtschaftswissenschaftlichen Fakultäten verschiedener deutscher Universitäten

An der *Westfälischen Wilhelms-Universität Münster* werden Lehrveranstaltungen der Wirtschaftswissenschaftlichen Fakultät mittels Austeilung von Papier-Fragebögen in den jeweiligen Veranstaltungen evaluiert. Der im Internet veröffentlichte Ablaufplan (vgl. o. V. 2012c) sieht dabei vor, dass die Evaluation gegen Ende der Veranstaltung erfolgen soll. Darüber hinaus ist der Evaluationszeitpunkt jedoch nicht näher spezifiziert und über den Tag der Befragung entscheidet jeder Dozent selbst. Die für die Evaluation benötigten Fragebögen werden bereits in den ersten Semesterwochen per E-Mail an die Dozenten gesendet, so dass diese hinsichtlich des Evaluationszeitpunkts durchaus Entscheidungsspielraum haben.

An der *Universität Regensburg* werden wie in Münster alle Lehrveranstaltungen der Wirtschaftswissenschaftlichen Fakultät einmal pro Semester mit einem papiergestützten Verfahren evaluiert. Die Fragebögen werden von studentischen Hilfskräften bzw. Assistenten des Studiendekans an die anwesenden Studierenden einer Lehrveranstaltung verteilt und anschließend wieder eingesammelt. Der Evaluationszeitpunkt ist dabei nicht vorgegeben, sondern die Dozenten können im Evaluationssystem der Universität Wunschtermine angeben (vgl. o. V. 2012a, S. 2). An der Fakultät für Betriebswirtschaftslehre der *Universität Mannheim* ist der Zeitpunkt der Evaluation in vergleichbarer Weise durch den Dozenten frei wählbar. Dieser verteilt zu einem selbst bestimmten Termin Papier-Fragebögen an seine Hörerschaft (vgl. o. V. 2012d).

Im Gegensatz zu den oben genannten Fakultäten führt der Fachbereich Wirtschaftswissenschaften der *Friedrich-Alexander Universität Erlangen-Nürnberg* Online-Evaluationen durch. Für die Durchführung der Evaluationen ist ein genauer Zeitplan vorgegeben (vgl. o. V. 2012b, S. 2). Dieser sieht in der Fassung von Mai 2012 vor, dass die Studierenden im Zeitraum vom 4. bis zum 29. Juni 2012 im Internet Bewertungen zu den von ihnen besuchten Lehrveranstaltungen abgeben können (vgl. o. V. 2012b, S. 2). Auch an der Wirtschaftswissenschaftlichen Fakultät der *Julius-Maximilians-Universität Würzburg* kommen Online-Befragungen zum Einsatz (wenngleich im Sommersemester 2012 die Mehrzahl der Lehrveranstaltungen mittels Papier-Fragebögen evaluiert wurde) (vgl. o. V. 2012e). Dabei ist die Online-Evaluation nicht erst am Veranstaltungsende vorgesehen, sondern die Möglichkeit zur Teilnahme endete im Sommersemester 2012 bereits am 26. Juni (vgl. o. V. 2012e).

Diese Beispiele zeigen, dass sich die Art und Weise, wie studentische Veranstaltungskritik eingeholt wird, teilweise zwischen den Fakultäten unterscheidet und dass – wie an der Julius-

Maximilians-Universität Würzburg – sogar innerhalb einer Fakultät für verschiedene Lehrveranstaltungen unterschiedliche Methoden zum Einsatz kommen können. Auch zeigt sich, dass keine einheitliche Regelung für den Evaluationszeitpunkt besteht. Zumindest an einigen Fakultäten kann dieser weitgehend frei vom jeweiligen Dozenten festgelegt werden.

3. Stand der Forschung

Studien, die Online-Evaluationen mit schriftlichen Befragungen in Lehrveranstaltungen vergleichen, weisen in der Regel das methodische Problem auf, dass es sich nicht um dieselben Teilnehmer handelt. So wurden zum Beispiel Lehrbewertungen verschiedener Semester verglichen (Avery/Bryant/Mathios/Kang/Bell 2006, S. 27) oder innerhalb einer Vorlesung wurden Gruppen von Studenten gebildet, denen divergierende Evaluations-Instrumentarien offeriert wurden (Layne/DeCristoforo/McGinty 1999, S. 224). Eine Ausnahme stellt in dieser Hinsicht eine Studie von Meinefeld dar, in der sich der Vergleich von Online- und Papier-Evaluation auf denselben Adressatenkreis bezieht, wobei beide Erhebungen wenige Wochen nacheinander durchgeführt wurden (vgl. Meinefeld 2010, S. 301). Nach eigenem Kenntnisstand gibt es bisher kaum entsprechende Studien (siehe auch Meinefeld 2010, S. 300).

Auch der Einfluss des Evaluationszeitpunkts wurde bisher nur vereinzelt untersucht. Laut Reichmann ist „bei (vergleichender) Analyse der Evaluationsergebnisse zu beachten, dass der Zeitpunkt der Evaluation durchaus Einfluss auf die Ergebnisse haben kann“ (Reichmann 2007, S. 79). Dieser Autor weist aber weder nach, dass ein solcher Einfluss besteht, noch erläutert er dessen tatsächliche oder vermutete Beschaffenheit. Eine Studie, die den Einfluss des Evaluationszeitpunkts tatsächlich untersucht, ist Pruitt/Dicks/Tilley (2011). Die für diese Studie befragten Studierenden gaben eine anonymen Kennung an, die sich zusammensetzt aus dem ersten Buchstaben ihrer Highschool, ihrem Geburtsmonat sowie den letzten beiden Ziffern ihrer Studentenidentifikationsnummer (vgl. Pruitt/Dicks/Tilley 2011, S. 7). Mit Hilfe dieser Kennung haben die Autoren 423 Paare von Evaluationsergebnissen verglichen, die sich auf 22 Lehrveranstaltungen beziehen, wobei jeweils zu Beginn und am Ende des Semesters evaluiert wurde (vgl. Pruitt/Dicks/Tilley 2011, S. 7). Bei Betrachtung aller Daten unterschieden sich die Evaluationszeitpunkte nicht signifikant (vgl. Pruitt/Dicks/Tilley 2011, S. 12). Allerdings zeigte sich, dass die aggregierte Betrachtung Details verschleiert: ‚Upper-Division-Studenten‘ tendierten dazu, ihre Dozenten am Semesterende schlechter zu bewerten, während bei ‚Lower-Division-Studenten‘ das Gegenteil beobachtbar war (vgl. Pruitt/Dicks/Tilley 2011, S. 12). Eine deutlich ältere Studie von Kohlan ergab, dass die Evaluations-

ergebnisse verschiedener Zeitpunkte hochgradig korrelieren und kein signifikanter Unterschied zwischen den Evaluationszeitpunkten besteht (vgl. Kohlan 1973, S. 589 und 591).

4. Methodik

4.1 Datenerhebung: mehrteilige Befragungen mit anonymen Kennungen

Die Datenerhebung erfolgte innerhalb eines Zeitraums von fünf Semestern (Wintersemester 2009/2010 bis Wintersemester 2011/2012) in 18 Wirtschaftsvorlesungen des Centrums für Interdisziplinäre Wirtschaftsforschung der Universität Münster, ehemals (bis September 2010) Institut für Ökonomische Bildung. Mit Ausnahme einer Vorlesung für den Masterstudiengang BWL richten sich die Lehrveranstaltungen des CIW vor allem an Studenten, die *kein* rein wirtschaftswissenschaftliches Fach (BWL oder VWL) studieren. Die Fächerkombinationen sind hier sehr vielfältig. Vertreten sind u. a. Studierende, die das Fach Wirtschaftswissenschaften mit dem Fach Politik, Jura, Erziehungswissenschaften, Sozialwissenschaften oder Theologie kombinieren, wobei ein Teil der Studierenden eine Lehramtslaufbahn anstrebt.

Im Rahmen einer Längsschnittstudie wurden die Studenten zu mehreren Semesterzeitpunkten (t) befragt. Evaluiert wurde rund einen Monat vor der Klausur ($t = 1$), in der letzten Vorlesungswoche ($t = 2$), kurz nach der Klausur ($t = 3$) und nach Notenbekanntgabe ($t = 4$). Alle Vorlesungen wurden mindestens einmal vor der Klausur (also in $t = 1$ und/oder in $t = 2$) sowie in $t = 3$ und in $t = 4$ evaluiert. In der Regel wurden dabei Online-Fragebögen verwendet (bei 57 von 60 Befragungen). Um die Antworten in den verschiedenen Befragungen den jeweiligen Studenten zuordnen zu können, wurden die Teilnehmer gebeten, bei jeder Befragung dieselbe, nur ihnen persönlich bekannte Kennung anzugeben. Die Kennung wurde so konzipiert, dass sie, wie die Matrikelnummer, individuell verschieden, im Gegensatz zu dieser aber anonym ist. In den ersten drei und im letzten Befragungssemester ergab sich die Kennung aus der Multiplikation des Geburtstags der Mutter (welcher z. B. der Zahl 104 entsprach, wenn diese am 1. April Geburtstag hatte) mit dem des Vaters. Studenten, denen einer dieser Geburtstage nicht bekannt war, wurden gebeten, ihn durch den Geburtstag einer anderen, aber in allen Befragungen identischen Person zu ersetzen. Im Sommersemester 2011 wurden nur zu den Zeitpunkten $t = 3$ und $t = 4$ Online-Befragungen durchgeführt. Vor dem Klausurtermin ($t = 2$) wurden, zwecks Generierung von Daten für ein weiteres Projekt, die offiziellen Papier-Fragebögen der Universität Münster um einen weiteren Fragebogen ergänzt. Um die Antworten auf den Papier-Fragebögen sowohl einander als auch den Ant-

worten in nachfolgenden Online-Fragebögen zuordnen zu können, wurden die Teilnehmer auch bei der Austeilung von Papier-Fragebögen gebeten, eine nur ihnen bekannte Kennung anzugeben. Allerdings musste die Formel für die anonyme Kennung für die Papier-Evaluationen und damit auch für die nachfolgenden Online-Evaluationen vereinfacht werden und entsprach dem Geburtsdatum der Mutter, also z. B. der Zahl 1041964, wenn diese am 1. April 1964 geboren wurde.

Insgesamt wurden 57 Online-Befragungen durchgeführt. In den ersten drei Befragungssemestern wurde das Online-Fragebogensystem *maQ* verwendet, das im Internet zur kostenlosen Nutzung zur Verfügung steht. Im Sommersemester 2011 und im Wintersemester 2011/2012 wurde das System *EvaSys* verwendet, das auch für die offiziellen (je nach Fachbereich papier- oder online-basierten) Lehrevaluationen der Universität Münster genutzt wird.² Der Link und die Zugangsdaten zur Befragung wurden per E-Mail an alle Studenten versendet, die aufgrund ihrer Klausuranmeldung oder ihrer Anmeldung auf Online-Lernplattformen – auf denen sich Studenten in einigen Vorlesungen anmelden konnten, um Materialien zu beziehen – mutmaßliche Teilnehmer der betreffenden Vorlesung waren.

4.2 Offizielle Studentische Lehrevaluationen mittels Papier-Fragebögen

Für alle Vorlesungen, für die Ergebnisse von Online-Evaluationen für den Zeitpunkt vorliegen, in dem auch die offiziellen Lehrevaluationen durchzuführen sind (letzte Vorlesungswoche), wurden auch die Daten der offiziellen papier-gestützten Evaluationen verwendet. Die offiziellen studentischen Lehrevaluationen der Wirtschaftswissenschaftlichen Fakultät der Westfälischen Wilhelms-Universität Münster wurden mittels Austeilung von Papier-Fragebögen in Lehrveranstaltungen durchgeführt. Die Ergebnisse dieser Evaluationen wurden vom Evaluationskoordinator bzw. der Evaluationskoordinatorin der Wirtschaftswissenschaftlichen Fakultät in Form von SPSS-Tabellen zur Verfügung gestellt. Jede Zeile einer SPSS-Tabelle entsprach dabei einem ausgefüllten Evaluationsbogen, d. h. der Bewertung einer bestimmten Lehrveranstaltung durch jeweils einen (anonymen) Studenten.

4.3 Verwendete Daten, Datenaufbereitung und betrachtete Ergebnisgrößen

Auf der Basis der eigenen Daten und der Ergebnisse der offiziellen Evaluationen wurden zwei Datensätze erstellt. *Datensatz I* beinhaltet ausschließlich Ergebnisse von Befragungen in der

² EvaSys (<https://zivevasys.uni-muenster.de/evasy/indexeva.php>) hat gegenüber maQ (<http://maq-online.de>) den Vorteil, dass Mehrfachantworten ausgeschlossen werden können, indem das System jedem Befragungsteilnehmer ein automatisch generiertes Passwort zusendet, welches für jeden Teilnehmer verschieden und nur dem System bekannt ist. Außerdem erlaubt EvaSys im Gegensatz zu maQ das Erstellen von Fragebogenkopien, die als Basis für die Erstellung neuer Fragebögen genutzt werden können.

letzten Vorlesungswoche. Um eine Vergleichbarkeit mit den Online-Befragungen zu erreichen, wurde bei den Papier-Evaluationen aus jeweils mehreren Abfragen zu den Bereichen «Struktur der Veranstaltung», «Engagement des Lehrenden», «Vermittlung der Lehrinhalte» sowie «eingesetzte Medien und zur Verfügung gestelltes Material» ein arithmetischer Mittelwert berechnet.³ In den Online-Befragungen wurden diese Bereiche jeweils in einer Frage abgehandelt, um die einzelnen Befragungen nicht zu umfangreich werden zu lassen, da dieselben Studenten Vorlesungen mehrfach bewerten sollten. Datensatz I wurde für die Untersuchung des Einflusses der Erhebungsart erstellt und beinhaltet Bewertungen von 14 Vorlesungen mit insgesamt 1.204 Fällen. Ein Fall entspricht dabei einem ausgefüllten Fragebogen. Studenten, die an mehreren Befragungen teilgenommen haben, stellen entsprechend der Anzahl an ausgefüllten Fragebögen mehrere Fälle dar. Von den 1.204 Fällen entstammen 1.003 Papier- und 201 Online-Evaluationen.

Datensatz II wurde erstellt, um in den Ergebnissen der Online-Befragungen nach Hinweisen auf einen systematischen Einfluss des Evaluationszeitpunkts zu suchen. In *Datensatz II* liegen 771 Fälle vor. Ein Fall entspricht dabei einem Studenten, der in einem Semester eine Vorlesung mindestens einmal bewertet hat.⁴ Studenten, die mehrere Vorlesungen evaluiert haben, stellen dementsprechend mehrere Fälle dar. Für Fragebögen mit identischer Kennung wurde angenommen, dass sie von demselben Studenten stammen, da die Wahrscheinlichkeit einer zufallsbedingten Übereinstimmung sehr gering ist.⁵ Demgemäß wurden die Antworten kennungsgleicher Fragebögen zu einer in einem bestimmten Semester gehaltenen Lehrveranstaltung in einer Zeile des Datensatzes zusammengefasst.

Um zu untersuchen, ob die Erhebungsart und der Evaluationszeitpunkt Einfluss auf studentische Evaluationen haben, wurden mehrere Ergebnisgrößen betrachtet: Rücklaufquoten, Gesamt- und Teilbereichsbewertungen sowie die Standardabweichungen der Teilbereichsbewertungen. Die Bewertungen beziehen sich auf eine fünfstufige, am deutschen Schulnotensystem orientierte Skala. Die Standardabweichung der Teilbereichsbewertungen wurde als Maß dafür betrachtet, wie differenziert Studenten Vorlesungen bewertet haben. Die Idee, die Ergebnisqualität von Evaluationen in dieser Weise zu operationalisieren, basiert auf der Beobachtung, dass die Teilbereichsbewertungen korreliert sind. Spearman Rangkorrelationen, be-

³ Die im Intranet der Westfälischen Wilhelms-Universität Münster veröffentlichten Ergebnisse der studentischen Veranstaltungskritik weisen für diese Bereiche ebenfalls ausschließlich Mittelwerte aus.

⁴ Bei Studenten, die keine Kennung angegeben haben, stellt jeder ausgefüllte Fragebogen einen Fall dar.

⁵ Doppelte Kennungen für Bewertungen derselben Vorlesungen zum gleichen Zeitpunkt traten nur in Fällen auf, in denen die angegebene Berechnungsformel offensichtlich nicht angewendet worden war (z. B. Kennungen mit einer zu geringen Anzahl von Ziffern, aus Buchstaben bestehende Kennungen oder Kennungen mit einer Abfolge identischer Zahlen). Solche Antworten wurden ohne Kennung in den Datensatz eingepflegt.

rechnet für alle Kombinationen von Teilbereichsbewertungen, sind für die Ergebnisse beider Erhebungsarten und aller Befragungszeitpunkte auf dem 0,1 %-Niveau signifikant. Vor diesem Hintergrund kann beim Vergleich von Online- und Papier-Evaluationen sowie bei Vergleichen von zwei Evaluationszeitpunkten eine größere Standardabweichung als ein Indiz für eine differenziertere Bewertung seitens der Studierenden betrachtet werden. Für die schlechter differenzierende Vergleichsgruppe kann vermutet werden, dass der Gesamteindruck bzw. die Wahrnehmung eines Teilaspekts der Vorlesung im Sinne eines Halo-Effekts (siehe z. B. Clayson/Haley 2011, S. 102) stärker auf die übrigen Bewertungen abgefärbt hat.

4.4 Überprüfung des Einflusses von Erhebungsart und -zeitpunkt

Um den Einfluss der *Erhebungsart* auf studentische Evaluationen zu untersuchen, wurde für einzelne Vorlesungen ($N = 14$), für die auf denselben Evaluationszeitpunkt bezogene Ergebnisse von Online- und Papier-Evaluationen vorliegen, mit Hilfe von Wilcoxon-Tests geprüft, ob sich die Ergebnisse signifikant unterscheiden. Um den Einfluss des *Evaluationszeitpunkts* zu untersuchen, wurden jeweils die Evaluationsergebnisse zweier Evaluationszeitpunkte verglichen und Wilcoxon-Tests berechnet. Dabei wurden zwei methodische Varianten betrachtet. Im Falle der methodischen *Variante 1* wird akzeptiert, dass beim Vergleich von zu zwei Zeitpunkten erhobenen Evaluationsergebnissen die Zusammensetzung der Teilnehmer variiert. In einer vorlesungsübergreifenden Betrachtung werden nur die Mittelwerte der Evaluationsergebnisse für die betreffenden Vorlesungen berücksichtigt und die statistischen Tests werden auf die Differenz der Mittelwerte für die verglichenen Zeitpunkte angewandt. Bei dieser Vorgehensweise werden alle Vorlesungen gleich gewichtet, unabhängig von ihrem Anteil an den ausgefüllten Fragebögen.

Allerdings verbleibt dabei die Möglichkeit, dass sich Selbstselektionseffekte bemerkbar machen, die durch die betrachteten Evaluationszeitpunkte bedingt sind. Nach Notenbekanntgabe könnten sich z. B. vornehmlich Studenten beteiligen, die von ihrer Note enttäuscht sind und nun durch entsprechende Lehrbewertungen ihren Unmut zum Ausdruck bringen. Methodische *Variante 2* ist daher so konstruiert, dass Selbstselektionseffekten ausgeschlossen werden. Dazu wurden bei Vergleichen von zwei Evaluationszeitpunkten nur Angaben von Studenten betrachtet, die die Vorlesung zu beiden Zeitpunkten bewertet haben.

5. Empirische Ergebnisse

5.1 Ergebnisse zeitgleicher Online- und Papier-Evaluationen

5.1.1 Rücklaufquoten von Online- und Papier-Evaluationen

Tabelle 1 zeigt, dass in allen 14 Vorlesungen deutlich weniger Studenten an der Online-Evaluation teilgenommen haben als an der kurz vor dem Klausurtermin ebenfalls durchgeführten offiziellen Lehrevaluation, bei der Papier-Fragebögen verteilt wurden.

Vorlesung	N _{Papier}	N _{Online}	N _{Klausur} ⁶	R _{Papier} (in %)	R _{Online} (in %)
A	119	55	172	69,19	31,98
B	126	13	152	82,89	8,55
C	38	6	50	76,00	12,00
D	114	10	175	65,14	5,71
E	81	6	102	79,41	5,88
F	24	15	35	68,57	42,86
G	23	4	29	79,31	13,79
H	18	9	23	78,26	39,13
I ⁷	45	5	19	—	—
N	91	6	215	42,33	2,79
O	170	34	218	77,98	15,60
P	107	24	189	56,61	12,70
Q	17	8	25	68,00	32,00
R	30	6	32	93,75	18,75
Gesamt	1.003	201	1.436		
Durchschnitt				72,11	18,60
Rücklaufquote (in %) = (Teilnehmeranzahl an der Evaluation / Klausurteilnehmeranzahl) · 100					

Tabelle 1: Anzahl der Teilnehmer (N) an der Papier-Evaluation, der Online-Evaluation und der Klausur sowie daraus geschätzte Rücklaufquoten (R)

Berücksichtigt wurden 14 Vorlesungen; für die nicht betrachteten Vorlesungen J bis M liegen keine Online-Evaluationen für die letzte Vorlesungswoche vor.

⁶ Bei der Anzahl der Klausurteilnehmer sind Fehlversuche, d. h. Studenten, die trotz Anmeldung nicht zur Klausur erschienen sind, miteingerechnet.

⁷ Aufgrund prüfungsrechtlicher Besonderheiten (Einzelklausur für einen Teil der Studierenden im jeweiligen Semester und Modulabschlussklausur für den übrigen Teil der Studierenden im nachfolgenden Semester) kann für Vorlesung I keine Rücklaufquote berechnet werden.

Da die exakte Anzahl der Vorlesungsteilnehmer nicht bekannt (und über das Semester hinweg vermutlich auch variabel) ist, beziehen sich die Schätzungen der Rücklaufquoten in Tabelle 1 – in Ermangelung einer geeigneteren Größe – auf die Anzahl der Klausurteilnehmer. Für die in der letzten Vorlesungswoche durchgeführten Online-Evaluationen beträgt der Durchschnitt der geschätzten Rücklaufquoten 18,60 %, für die in der gleichen Woche durchgeführten Papier-Evaluationen hingegen 72,11 %.

Abweichungen zwischen der Anzahl der Klausurteilnehmer und der Anzahl ausgefüllter Papier-Fragebögen haben zumeist alltägliche Ursachen wie z. B. Abwesenheit zum Termin der Klausur oder der Evaluation, Verzicht auf Teilnahme an der Evaluation trotz Anwesenheit (z. B. wenn die Evaluationsbögen zum Schluss der jeweiligen Lehrveranstaltung verteilt werden) oder Klausurteilnahme ohne vorherigen Vorlesungsbesuch. Daneben können die Abweichungen aber auch spezieller Natur sein: Im Falle von Vorlesung I war die Anzahl der Teilnehmer an der offiziellen studentischen Lehrevaluation aufgrund von prüfungsrechtlichen Besonderheiten mehr als doppelt so groß wie die Anzahl der Klausurteilnehmer. Da es somit keine belastbare Zahl für die Anzahl der für die Evaluation teilnahmeberechtigten Studenten gab, musste hier auf eine Berechnung der Rücklaufquoten verzichtet werden.

5.1.2 Bewertungen einzelner Vorlesungen insgesamt und in Teilbereichen

Wenngleich sich die Rücklaufquoten von Online- und Papier-Evaluationen erheblich unterscheiden (siehe Tabelle 1), besteht ein deutlicher Zusammenhang zwischen den Ergebnissen, wie in Abbildung 1 gezeigt wird. Die Abszisse repräsentiert hier die durchschnittliche Bewertung in der Papier-Evaluation und die Ordinate die durchschnittliche Bewertung in der Online-Evaluation. Für jede Vorlesung werden fünf Bewertungen betrachtet, und zwar die Gesamtbewertung sowie vier Teilbereichsbewertungen. Bei perfekter Übereinstimmung der beiden Evaluationsarten liegt ein Datenpunkt auf der eingezeichneten Diagonale. Fälle, in denen die Online-Evaluation eine schlechtere Bewertung ergab als die Papier-Evaluation, sind oberhalb dieser Diagonale zu finden, gegenteilige Fälle darunter. Die Gesamtbewertungen werden durch ausgefüllte schwarze Kreise symbolisiert, während sich die nicht ausgefüllten Symbole auf die Teilbereichsbewertungen beziehen. Bei Letzteren führen die unterschiedlichen Erhebungsarten tendenziell zu stärkeren Abweichungen als bei der Gesamtbewertung, was in Abbildung 1 an den etwas größeren Abständen zur Diagonale erkennbar ist.

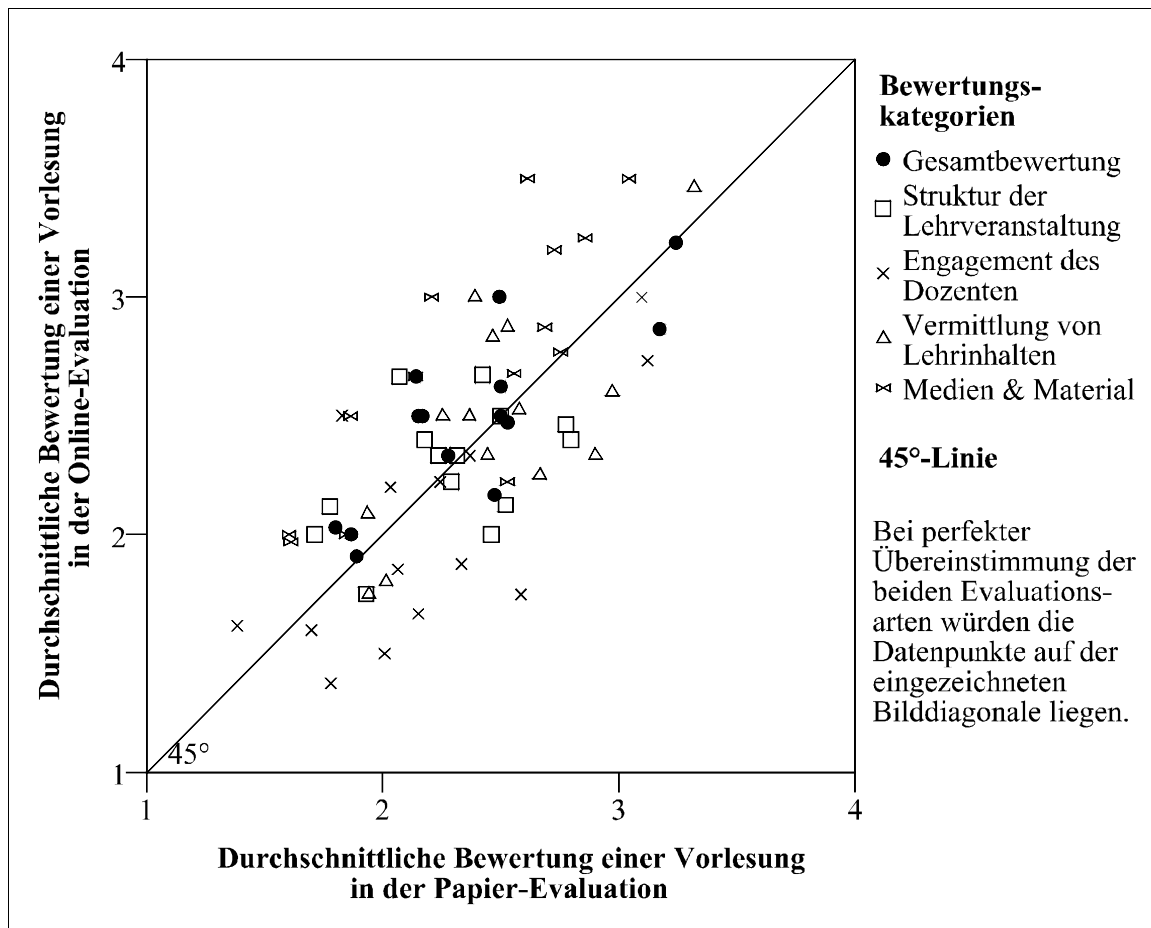


Abbildung 1: Vergleiche der durchschnittlichen Bewertungen bei Papier- und Online-Evaluation

		Papier-Evaluation				
		Gesamt	Struktur	Engagement	Vermittlung	Medien & Material
Online-Evaluation	Gesamt	0,81***				
	Struktur		0,44			
	Engagement			0,70**		
	Vermittlung				0,70**	
	Medien & Material					0,81***

Tabelle 2: Korrelation der durchschnittlichen Bewertungen einer Vorlesung bei Papier- und bei Online-Evaluation

Berechnung nach Pearson; $N = 14$; **/** kennzeichnet statistische Signifikanz auf dem 1/0,1 %-Niveau.

Basierend auf den in Abbildung 1 dargestellten Wertepaaren wurde für jede der fünf Bewertungskategorien eine Korrelation nach Pearson berechnet. Wie Tabelle 2 zeigt, sind die Ergebnisse von Online- und Papier-Evaluationen hochgradig korreliert, mit Ausnahme der Bewertung der Struktur der Vorlesungen (Struktur). Von dieser Ausnahme abgesehen, nehmen die Korrelationskoeffizienten Werte zwischen 0,70 und 0,81 an, was auf einen stark positiven Zusammenhang hinweist (vgl. Brosius 2011, S. 523).

Dass Vorlesungen bei Online-Evaluationen trotz deutlich geringerer Rücklaufquoten nicht völlig anders bewertet werden als bei Papier-Evaluationen, zeigen auch die in Tabelle 3 ausgewiesenen Differenzen der durchschnittlichen Gesamt- und Teilbereichsbewertungen. Ein negatives Vorzeichen bedeutet hier, dass die Bewertung in der Online-Evaluation schlechter ist als in der Papier-Evaluation. In 8 von 14 Vorlesungen unterscheiden sich die *Gesamtbewertungen* in Online- und Papier-Evaluation um weniger als 0,30 Notenpunkte; der maximale Unterschied beträgt 0,52 Notenpunkte. Hinsichtlich der *Teilbereichsbewertungen* unterscheiden sich die beiden Erhebungsarten in 27 von 56 Fällen um weniger als 0,30 Notenpunkte, wobei mit „Fall“ hier die durchschnittliche Bewertung einer Vorlesung in einem Teilbereich gemeint ist. In 10 Fällen (17,86 %) sind betragsmäßige Differenzen von 0,50 oder mehr Notenpunkten beobachtbar; der maximale Unterschied beträgt 0,89 Notenpunkte.

Tabelle 3 zeigt, dass 9 von 14 Vorlesungen in der Online-Evaluation eine schlechtere *Gesamtbewertung* erhielten als in der Papier-Evaluation; in einer Vorlesung wurde identisch evaluiert. Ein Wilcoxon-Test, angewandt auf die 14 Paare der durchschnittlichen Gesamtbewertungen (Gesamt), ergibt jedoch keinen signifikanten Unterschied zwischen Online- und Papier-Evaluationen. Wilcoxon-Tests für die durchschnittlichen Bewertungen der Vorlesungen in Teilbereichen ergeben nur in Bezug auf die Bewertungen von eingesetzten Medien und zur Verfügung gestelltem Material (Medien & Material) eine signifikante Abweichung zwischen beiden Erhebungsarten. Dieser Teilbereich wurde in den Online-Evaluationen im Durchschnitt um 0,36 Notenpunkte schlechter bewertet als in den Papier-Evaluationen. Für die Bewertungen des Engagements der Dozenten (Engagement) ergibt ein Wilcoxon-Test einen schwach signifikanten Unterschied. Dieser Aspekt wurde in den Online-Evaluationen im Durchschnitt um 0,18 Notenpunkte besser bewertet.

Tabelle 1 und Tabelle 3 zeigen, dass die Unterschiede zwischen Online- und Papier-Evaluation für die einzelnen Vorlesungen verschieden stark ausgeprägt sind, und zwar sowohl hinsichtlich der Rücklaufquoten als auch hinsichtlich der Lehrbewertungen. Zu klären ist, ob ein Zusammenhang zwischen diesen beiden Aspekten besteht. Der Unterschied in der Rück-

laufquote, der sich durch Abzug der Rücklaufquote der Online-Evaluation von der Rücklaufquote der Papier-Evaluation ergibt, ist für die 13 Vorlesungen, die hier berücksichtigt werden können, positiv.⁸ Berechnungen von Korrelationen nach Pearson ergeben jedoch, dass weder die betragsmäßigen Differenzen der Gesamtbewertungen noch die der einzelnen Teilbereichsbewertungen signifikant mit den Differenzen der Rücklaufquoten korrelieren.

		B _{Papier} - B _{Online}				
		Gesamt	Struktur	Engagement	Vermittlung	Medien & Material
Vorlesung	A	0,06	-0,25	0,21	0,05	-0,12
	B	0,01	0,32	0,10	-0,14	-0,02
	C	0,31	0,46	0,51	0,57	-0,79
	D	-0,33	-0,22	-0,17	-0,25	-0,46
	E	-0,35	-0,59	-0,67	-0,37	-0,64
	F	0,31	0,40	0,39	0,37	-0,47
	G	0,00	0,00	0,84	0,42	-0,39
	H	-0,06	0,07	0,02	0,11	0,31
	I	-0,13	-0,29	0,10	0,22	-0,40
	N	-0,51	-0,02	0,49	-0,61	-0,89
	O	-0,23	-0,34	-0,23	-0,15	-0,36
	P	-0,02	0,18	0,40	0,19	-0,15
	Q	-0,13	0,40	0,46	-0,35	-0,19
	R	-0,52	-0,10	0,04	-0,13	-0,53
Durchschnitt		-0,11	0,00	0,18	0,00	-0,36
Wilcoxon-Test	B _{Papier} < B _{Online}	9	7	3	7	13
	B _{Papier} > B _{Online}	4	6	11	7	1
	B _{Papier} = B _{Online}	1	1	0	0	0
	Signifikanz			(*)		**

Tabelle 3: Differenzen der durchschnittlichen Lehrbewertungen (B) von 14 Vorlesungen bei Papier- und bei Online-Evaluation

Schwarz markiert: Differenz $\geq 0,50$ Notenpunkte; grau markiert: $0,30$ Notenpunkte $\leq$ Differenz $< 0,50$ Notenpunkte; (*)/** kennzeichnet statistische Signifikanz auf dem 10/1 %-Niveau

⁸ Aufgrund prüfungsrechtlicher Besonderheiten kann für Vorlesung I keine Rücklaufquote berechnet werden.

5.1.3 Standardabweichung der Teilbereichsbewertungen

In 12 von 14 Vorlesungen ist die durchschnittliche Standardabweichung der Teilbereichsbewertungen bei der Online-Evaluation größer als bei der Papier-Evaluation. Der Unterschied beträgt im Durchschnitt 0,14 Notenpunkte. Mithilfe eines Wilcoxon-Tests lässt sich zeigen, dass sich beide Erhebungsarten in Bezug auf die Standardabweichung der Teilbereichsbewertungen signifikant unterscheiden (das statistische Signifikanzniveau beträgt 1 %). Dies deutet darauf hin, dass Teilnehmer an Online-Evaluationen differenziertere Bewertungen abgeben als Teilnehmer an Papier-Evaluationen.

5.2 Ergebnisse von Online-Evaluationen zu verschiedenen Zeitpunkten

5.2.1 Durchschnittliche Vorlesungsbewertungen und Rücklaufquoten

Für Untersuchungen eines möglicherweise bestehenden Einflusses des Evaluationszeitpunkts stehen neben den 14 Vorlesungen, die oben für den Vergleich von Online- und Papier-Evaluationen herangezogen wurden, vier weitere Vorlesungen zur Verfügung. Allerdings wurden nicht alle Vorlesungen zu vier Zeitpunkten evaluiert und in einigen Fällen fanden keine Online-Evaluationen statt, sondern stattdessen papierbasierte Evaluationen. Von daher variiert bei paarweisem Vergleich aller Kombinationen von Evaluationszeitpunkten die Anzahl der zur Verfügung stehenden Lehrevaluationen. Tabelle 4 ist zu entnehmen, zu welchen Zeitpunkten ($1 \leq t \leq 4$) in den mit A bis R bezeichneten Vorlesungen Online-Evaluationen stattfanden und wie viele Studenten hieran teilgenommen haben (N_{Online}). Unter Berücksichtigung der Teilnehmerzahl an der Klausur (N_{Klausur}) wurden Rücklaufquoten (R_{Online}) berechnet.

Für die Gesamtheit aller Vorlesungen ergeben sich insgesamt 63 Möglichkeiten, verschiedene Evaluationszeitpunkte paarweise zu vergleichen. In Abbildung 2 ist jede dieser Vergleichsmöglichkeiten als Punkt dargestellt, dessen Lage sich aus der mittleren Gesamtbewertung zu den beiden Zeitpunkten ergibt. Dabei ist der frühere Evaluationszeitpunkt (t_0) der Abszisse zugeordnet und der spätere Evaluationszeitpunkt (t_1) der Ordinate. Wie stark sich die durchschnittlichen Gesamtbewertungen zu zwei Evaluationszeitpunkten unterscheiden, ist am Abstand des betreffenden Datenpunktes von der eingezeichneten Diagonale erkennbar, die einer perfekten Übereinstimmung beider Evaluationsergebnisse entspricht. Oberhalb der Diagonale sind alle Fälle angesiedelt, in denen eine Vorlesung zu einem früheren Evaluationszeitpunkt besser bewertet wurde; gegenteilige Fälle sind unterhalb der Diagonale positioniert.

Vorlesung	N _{Klausur}	N _{Online}				R _{Online} (in %)			
		<i>t</i> = 1	<i>t</i> = 2	<i>t</i> = 3	<i>t</i> = 4	<i>t</i> = 1	<i>t</i> = 2	<i>t</i> = 3	<i>t</i> = 4
A	172		55	43	39		31,98	25,00	22,67
B	152	22	13	12	8	14,47	8,55	7,89	5,26
C	50	0	6	6	2	0,00	12,00	12,00	4,00
D	175	17	10	6	9	9,71	5,71	3,43	5,14
E	102	11	6	5	7	10,78	5,88	4,90	6,86
F	35		15	12	11		42,86	34,29	31,43
G	29	4	4	3	3	13,79	13,79	10,34	10,34
H	23		9	8	9		39,13	34,78	39,13
I ⁹	19	12	5	7	6	—	—	—	—
J	206	34		41	35	16,50		19,90	16,99
K	142			39	22			27,46	15,49
L	41			23	11			56,10	26,83
M	15			2	2			13,33	13,33
N	215		6	20	22		2,79	9,30	10,23
O	218		34	25	16		15,60	11,47	7,34
P	189		24	18	21		12,70	9,52	11,11
Q	25		8	7	5		32,00	28,00	20,00
R	32		6	4	4		18,75	12,50	12,50
Gesamt	1840	100	201	281	232				
Durchschnitt						10,88	18,60	18,84	15,21
Rücklaufquote (in %) = (Teilnehmeranzahl an der Evaluation / Klausurteilnehmeranzahl)·100									

Tabelle 4: Anzahl der Teilnehmer (N) an der Klausur bzw. den zu verschiedenen Zeitpunkten (t) erfolgten Online-Evaluationen und daraus berechnete Rücklaufquoten (R)
 Evaluiert wurde ca. einen Monat vor der Klausur (t = 1), in der letzten Vorlesungswoche (t = 2), nach der Klausur (t = 3) und nach Notenbekanntgabe (t = 4).

Abbildung 2 zeigt, dass zwischen den Gesamtbewertungen von Vorlesungen zu zwei Zeitpunkten ein deutlicher Zusammenhang besteht. Berechnungen von Korrelationen nach Pearson, deren Ergebnisse in Tabelle 5 dargestellt sind, sind für alle sechs Zeitpunktvergleiche signifikant. Die Korrelationskoeffizienten nehmen Werte zwischen 0,86 und 0,96 an. Dies lässt auf einen starken positiven Zusammenhang zwischen den Gesamtbewertungen von Vorlesungen zu verschiedenen Zeitpunkten schließen.

⁹ Aufgrund prüfungsrechtlicher Besonderheiten kann für Vorlesung I keine Rücklaufquote berechnet werden.

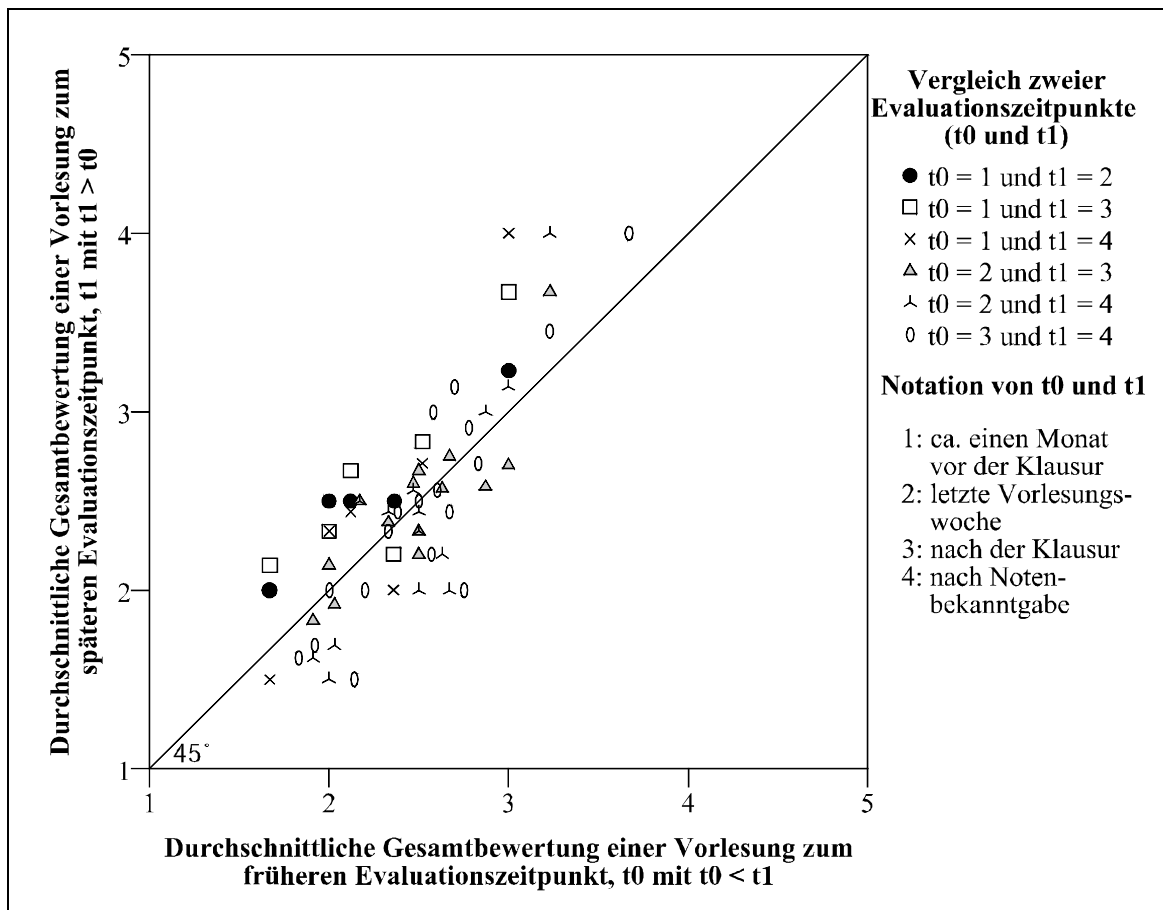


Abbildung 2: Paarweiser Vergleich der durchschnittlichen Gesamtbewertungen von Vorlesungen zu verschiedenen Evaluationszeitpunkten

		Gesamtbewertung zum früheren Zeitpunkt (t_0)		
		$t_0 = 1$	$t_0 = 2$	$t_0 = 3$
Gesamtbewertung zum späteren Zeitpunkt (t_1)	$t_1 = 2$	0,96** (N = 5)		
	$t_1 = 3$	0,87* (N = 6)	0,86*** (N = 14)	
	$t_1 = 4$	0,90* (N = 6)	0,86*** (N = 14)	0,89*** (N=18)

Tabelle 5: Korrelation der durchschnittlichen Gesamtbewertungen von Vorlesungen zu verschiedenen Evaluationszeitpunkten

Evaluiert wurde ca. einen Monat vor der Klausur ($t = 1$), in der letzten Vorlesungswoche ($t = 2$), nach der Klausur ($t = 3$) und nach Notenbekanntgabe ($t = 4$); Berechnung nach Pearson; */**/** kennzeichnet statistische Signifikanz auf dem 5/1/0,10 %-Niveau.

Für alle Kombinationen der vier Evaluationszeitpunkte wurde außerdem die mittlere Differenz der Gesamt- und der Teilbereichsbewertungen berechnet und mit Wilcoxon-Tests wurde geprüft, ob signifikante systematische Unterschiede zwischen den Zeitpunkten bestehen. Die Ergebnisse sind in Tabelle 6 dargestellt. Allen Vergleichen, bei denen der Wilcoxon-Test ein signifikantes Ergebnis liefert, ist gemein, dass die mittlere Differenz der Bewertungen negativ ist. Dies bedeutet, dass zum früheren Zeitpunkt (t_0) im Durchschnitt besser bewertet wurde (es wurde jeweils die spätere Bewertung von der früheren subtrahiert). Der Vergleich der Evaluationszeitpunkte 1 (ca. einen Monat vor der Klausur) und 2 (letzte Vorlesungswoche) ergibt signifikante Unterschiede für die Gesamtbewertung (Gesamt) sowie die Bereiche «Struktur» und «Medien & Material»; für den Bereich «Vermittlung» ist der Unterschied schwach signifikant. Zum ersten Evaluationszeitpunkt wurden Vorlesungen in diesen Bereichen im Durchschnitt um 0,14 bis 0,38 Notenpunkte besser bewertet als zum zweiten. Die Evaluationszeitpunkte 1 und 3 (nach der Klausur) unterscheiden sich signifikant bezüglich der Gesamtbewertung sowie der Bewertung in den Bereichen «Struktur» und «Medien & Material»; schwach signifikant ist der Unterschied im Bereich «Engagement». Die mittleren Differenzen betragen je nach Bewertungskategorie 0,35 bis 0,51 Notenpunkte. Für die Evaluationszeitpunkte 1 und 4 (nach Notenbekanntgabe) zeigen sich signifikante Unterschiede in den Bereichen «Struktur» und «Engagement»; zum Evaluationszeitpunkt 1 wurden Vorlesungen in diesen Bereichen im Durchschnitt um 0,41 bzw. 0,49 Notenpunkte besser bewertet als zum Evaluationszeitpunkt 4. Der zweite und der dritte Zeitpunkt unterscheiden sich signifikant hinsichtlich der «Struktur» und schwach signifikant hinsichtlich des «Engagements»; die mittleren Differenzen betragen jeweils 0,17 Notenpunkte. Vergleiche der Evaluationszeitpunkte 2 und 4 ergeben nur für den Bereich «Struktur» einen schwach signifikanten Unterschied. Die Evaluationszeitpunkte 3 und 4 unterscheiden sich in keiner Bewertungskategorie signifikant.

Wilcoxon-Tests für die *Standardabweichungen der Teilbereichsbewertungen* ergeben lediglich für den Vergleich des zweiten mit dem dritten Evaluationszeitpunkt einen schwach signifikanten Unterschied (zum zweiten Zeitpunkt fällt die Standardabweichung bei 10 von 14 Vorlesungen höher aus). Für die übrigen Vergleiche zweier Evaluationszeitpunkte gibt es somit keinen Anhaltspunkt für einen Einfluss des Evaluationszeitpunkts auf die Differenziertheit der Lehrbewertungen. Wilcoxon-Tests für die *Rücklaufquoten* der Online-Evaluationen (siehe Tabelle 4) ergeben signifikante Unterschiede für Vergleiche der Evaluationszeitpunkte 2 und 3, 2 und 4 sowie 3 und 4. Wie Tabelle 7 zeigt, ist hierbei in der Mehrzahl der Vorlesungen der Rücklauf zum früheren Evaluationszeitpunkt (t_0) höher als zum späteren Zeitpunkt (t_1).

Evaluationszeitpunkte (t_0 und t_1)	Bewertungskategorien	N (Vorlesungen)	Mittlere Differenz ($B_{t_0} - B_{t_1}$)	Wilcoxon-Test			
				$B_{t_0} < B_{t_1}$	$B_{t_0} > B_{t_1}$	$B_{t_0} = B_{t_1}$	Signifikanz
$t_0 = 1$ und $t_1 = 2$	Gesamt	5	-0,32	5	0	0	*
	Struktur	5	-0,38	5	0	0	*
	Engagement	5	-0,23	4	1	0	
	Vermittlung	5	-0,14	4	0	1	(*)
	Medien & Material	5	-0,37	5	0	0	*
$t_0 = 1$ und $t_1 = 3$	Gesamt	6	-0,36	5	1	0	*
	Struktur	6	-0,51	6	0	0	*
	Engagement	6	-0,35	5	1	0	(*)
	Vermittlung	6	-0,23	4	1	1	
	Medien & Material	6	-0,47	5	1	0	*
$t_0 = 1$ und $t_1 = 4$	Gesamt	6	-0,22	4	2	0	
	Struktur	6	-0,41	6	0	0	*
	Engagement	6	-0,49	5	0	1	*
	Vermittlung	6	-0,10	4	1	1	
	Medien & Material	6	-0,26	3	1	2	
$t_0 = 2$ und $t_1 = 3$	Gesamt	14	0,00	7	7	0	
	Struktur	14	-0,17	10	3	1	*
	Engagement	14	-0,17	11	3	0	(*)
	Vermittlung	14	-0,10	10	4	0	
	Medien & Material	14	-0,05	7	7	0	
$t_0 = 2$ und $t_1 = 4$	Gesamt	14	0,10	6	8	0	
	Struktur	14	-0,20	10	3	1	(*)
	Engagement	14	-0,13	8	6	0	
	Vermittlung	14	0,05	6	7	1	
	Medien & Material	14	0,06	5	8	1	
$t_0 = 3$ und $t_1 = 4$	Gesamt	18	0,07	6	9	3	
	Struktur	18	-0,04	7	9	2	
	Engagement	18	0,03	11	6	1	
	Vermittlung	18	0,10	8	9	1	
	Medien & Material	18	0,04	6	10	2	

Tabelle 6: Vergleiche der durchschnittlichen Bewertungen (B) derselben Vorlesungen zu verschiedenen Evaluationszeitpunkten

Evaluert wurde ca. einen Monat vor der Klausur ($t = 1$), in der letzten Vorlesungswoche ($t = 2$), nach der Klausur ($t = 3$) und nach Notenbekanntgabe ($t = 4$); (*)/* kennzeichnet statistische Signifikanz auf dem 10/5 %-Niveau.

Evaluationszeitpunkte (t_0 und t_1)	N (Vorlesungen)	Mittlere Differenz ($R_{t_0} - R_{t_1}$)	Wilcoxon-Test			
			$R_{t_0} < R_{t_1}$	$R_{t_0} > R_{t_1}$	$R_{t_0} = R_{t_1}$	Signifikanz
$t_0 = 1$ und $t_1 = 2$	5	0,56	1	3	1	
$t_0 = 1$ und $t_1 = 3$	6	1,13	2	4	0	
$t_0 = 1$ und $t_1 = 4$	6	2,78	2	4	0	
$t_0 = 2$ und $t_1 = 3$	13	2,95	1	11	1	*
$t_0 = 2$ und $t_1 = 4$	13	4,29	2	10	1	*
$t_0 = 3$ und $t_1 = 4$	17	3,62	5	9	3	*

Tabelle 7: Vergleiche der Rücklaufquoten (R) von Online-Evaluationen zu verschiedenen Evaluationszeitpunkten

Evaluert wurde ca. einen Monat vor der Klausur ($t = 1$), in der letzten Vorlesungswoche ($t = 2$), nach der Klausur ($t = 3$) und nach Notenbekanntgabe ($t = 4$); * kennzeichnet statistische Signifikanz auf dem 5 %-Niveau.

5.2.2 Lehrbewertungen derselben Studenten zu verschiedenen Evaluationszeitpunkten

Tabelle 8 zeigt, dass Unterschiede zwischen den Evaluationszeitpunkten auch bei Mehrfachbewertungen durch *dieselben* Studenten zu finden sind. Eine solche Betrachtung schließt Selbstselektionseffekte, die durch den Evaluationszeitpunkt ausgelöst werden könnten, aus. Wie bei der obigen Betrachtung der Vorlesungsmittelwerte zeigen sich signifikante Unterschiede ausnahmslos in Form einer besseren Bewertung des früheren Zeitpunkts (t_0). Bei Vergleichen der Evaluationszeitpunkte 1 und 2 ergeben Wilcoxon-Tests nur in Bezug auf den Bereich «Struktur» einen schwach signifikanten Unterschied; die Evaluationszeitpunkte 1 und 3 unterscheiden sich hinsichtlich keiner der betrachteten Lehrbewertungskategorien. Dies deutet darauf hin, dass die für dieselben Zeitpunktvergleiche erhaltenen Ergebnisse in Tabelle 6, die ohne Ausschluss von möglichen Selbstselektionseffekten zustande kamen, vor allem Resultat einer variierenden Zusammensetzung der Befragungsteilnehmer sind.

Für Vergleiche der Evaluationszeitpunkte 1 und 4 zeigen sich hingegen auch unter Ausschluss von Selbstselektionseffekten signifikante Unterschiede bezüglich der Bewertungen in den Bereichen «Struktur» und «Engagement»; daneben unterscheiden sich die beiden Evaluationszeitpunkte schwach signifikant im Bereich «Medien und Material».

Evaluations- zeitpunkte (t_0 und t_1)	Bewertungs- kategorien	N (Stu- denten)	Mittlere Differenz ($B_{t_0} - B_{t_1}$)	Wilcoxon-Test			
				$B_{t_0} < B_{t_1}$	$B_{t_0} > B_{t_1}$	$B_{t_0} = B_{t_1}$	Signi- fikanz
$t_0 = 1$ und $t_1 = 2$	Gesamt	26	-0,08	8	6	12	
	Struktur	26	-0,19	7	2	17	(*)
	Engagement	26	-0,12	7	4	15	
	Vermittlung	26	-0,04	5	4	17	
	Medien & Material	26	0,08	7	9	10	
$t_0 = 1$ und $t_1 = 3$	Gesamt	37	-0,05	8	6	23	
	Struktur	37	-0,14	11	6	20	
	Engagement	36	-0,19	10	7	19	
	Vermittlung	37	-0,03	9	7	21	
	Medien & Material	36	-0,11	12	11	13	
$t_0 = 1$ und $t_1 = 4$	Gesamt	37	-0,05	10	7	20	
	Struktur	37	-0,22	12	4	21	*
	Engagement	36	-0,39	17	6	13	*
	Vermittlung	37	-0,19	10	5	22	
	Medien & Material	37	-0,27	14	8	15	(*)
$t_0 = 2$ und $t_1 = 3$	Gesamt	112	-0,04	17	14	81	
	Struktur	113	-0,07	29	19	65	
	Engagement	113	-0,12	24	11	78	*
	Vermittlung	110	-0,10	30	20	60	
	Medien & Material	111	-0,23	31	11	69	***
$t_0 = 2$ und $t_1 = 4$	Gesamt	86	-0,07	20	15	51	
	Struktur	88	-0,07	21	15	52	
	Engagement	88	-0,20	27	12	49	**
	Vermittlung	87	-0,17	32	18	37	(*)
	Medien & Material	86	-0,22	27	12	47	*
$t_0 = 3$ und $t_1 = 4$	Gesamt	133	0,01	21	22	90	
	Struktur	134	0,00	24	25	85	
	Engagement	133	-0,18	35	16	82	**
	Vermittlung	130	0,00	27	28	75	
	Medien & Material	133	-0,04	30	29	74	

Tabelle 8: Vergleiche der Lehrbewertungen (B) derselben Studenten zu verschiedenen Evaluationszeitpunkten

Evaluert wurde ca. einen Monat vor der Klausur ($t = 1$), in der letzten Vorlesungswoche ($t = 2$), nach der Klausur ($t = 3$) und nach Notenbekanntgabe ($t = 4$); (*)/**/***/

kennzeichnet statistische Signifikanz auf dem 10/5/1/0,1 %-Niveau

Der zweite und der dritte Evaluationszeitpunkt unterscheiden sich signifikant hinsichtlich der Bewertungen in den Bereichen «Engagement» und «Medien und Material». Gleiches gilt für die Evaluationszeitpunkte 2 und 4, die sich zudem auch schwach signifikant bezüglich der Bewertungen im Bereich «Vermittlung» unterscheiden. Ein Vergleich der Evaluationszeitpunkte 3 und 4 ergibt lediglich für den Bereich «Engagement» einen signifikanten Unterschied. Die durchschnittlichen Gesamtbewertungen zeigen unter Ausschluss von Selbstselektionseffekten in keinem der sechs Vergleiche zweier Evaluationszeitpunkte signifikante Unterschiede. Für die Standardabweichungen der Teilbereichsbewertungen ergeben Berechnungen von Wilcoxon-Tests lediglich für einen Vergleich der Evaluationszeitpunkte 2 und 3 einen signifikanten Unterschied, der auf differenziertere Bewertungen zum früheren Evaluationszeitpunkt hindeutet.

6. Diskussion

6.1 Erhebungsart

Sowohl für die eigenen Ergebnisse als auch die von Meinefeld (2010) lässt sich nicht ausschließen, dass die geringen Rücklaufquoten bei den Online-Evaluationen zumindest teilweise eine unvermeidbare Konsequenz des Studiendesigns sind. Dieses setzt die Bereitschaft von Studenten voraus, ein und dieselbe Lehrveranstaltung mehrfach zu bewerten. Die Vermutung, dass die unterschiedlichen Rücklaufquoten nur partiell einen prinzipiellen Unterschied zwischen Online- und Papier-Evaluationen widerspiegeln, wird dadurch erhärtet, dass die Unterschiede in anderen Untersuchungen, die nicht mit einem erhöhten Aufwand für die Studenten verbunden waren, deutlich geringer ausfallen. So beträgt der auf die Durchschnittswerte bezogene Unterschied der Rücklaufquoten bei Layne/DeChristoforo/McGinty (1999, S. 26) nur 12,8 %, ¹⁰ und auch in der Studie von Avery/Bryant/Mathios/Kang/Bell (2006, S. 27) sind die Unterschiede kleiner ¹¹. Letztere berichten sogar, dass nach Umstellung auf Online-Befragungen die Rücklaufquoten stiegen (vgl. Avery/Bryant/Mathios/Kang/Bell 2006, S. 36).

¹⁰ In den papiergestützten Evaluationen betrug die Rücklaufquote im Durchschnitt 60,6 % und in den Online-Evaluationen 47,8 % (vgl. Layne/DeChristoforo/McGinty 1999, S. 26).

¹¹ Im Durchschnitt beträgt die Rücklaufquote aller Papier-Evaluationen (N = 20) – nach eigener Berechnung, basierend auf den veröffentlichten Rücklaufquoten der einzelnen Lehrveranstaltungen – 77,8 %; bei den Online-Evaluationen (N = 9) ergibt sich ein arithmetisches Mittel von 49,8 %. In den beiden Fällen, in denen ein Professor in einem Semester die gleiche Vorlesung zu zwei Terminen hielt, wobei für den einen Termin Papier- und für den anderen Termin Online-Evaluationen durchgeführt wurden, beträgt die Rücklaufquote der Papier-Evaluation 69,2 % (im Jahr 2000) bzw. 72,9 % (im Jahr 2001) und die Rücklaufquote der Online-Evaluation 46,7 % (im Jahr 2000) bzw. 48,5 % (im Jahr 2001) (vgl. Avery/Bryant/Mathios/Kang/Bell 2006, S. 28 und 29).

Die Ergebnisse von Meinefeld geben keinen Hinweis darauf, dass die Erhebungsart einen signifikanten systematischen Einfluss auf die Bewertungen von Vorlesungen hat (vgl. Meinefeld 2010, S. 305). Für die eigenen Ergebnisse zeigten sich signifikante systematische Unterschiede nur für die Bewertungen im Bereich «*Medien & Material*»; schwach signifikante Unterschiede zeigten sich hinsichtlich der Bewertungen im Bereich «*Engagement*». Demnach können zumindest die mit unterschiedlichen Erhebungsarten erhaltenen *Gesamtbewertungen* von Vorlesungen durchaus sinnvoll miteinander verglichen werden. Wenngleich hierbei nicht mit größeren *systematischen* Unterschieden zu rechnen ist, können einzelne Bewertungen aber deutlich voneinander abweichen, wie sowohl die Ergebnisse von Meinefeld als auch die eigenen Ergebnissen zeigen (siehe hier Tabelle 3 und Meinefeld 2010, S. 306 f.).

Aus den Bewertungsdifferenzen zieht Meinefeld folgenden Schluss: „Eine Online-Evaluation ist zwar kostengünstig und schnell, aber ihre Ergebnisse sind unzuverlässig, da die Veranstaltungen durchweg falsch bewertet werden.“ (Meinefeld 2010, S. 308). Dieser Schluss erscheint aber in mehrfacher Hinsicht diskussionsbedürftig. So geben Abweichungen der Ergebnisse beider Erhebungsmethoden keinen Hinweis darauf, welche Methode die Lehrqualität zutreffender erfasst. Die Interpretation von Meinefeld (2010, S. 302), der zufolge etwaige Ergebnisdifferenzen auf Defizite der Online-Evaluationen zurückzuführen sind, wird damit begründet, dass bei schriftlichen Befragungen in Lehrveranstaltungen die Rücklaufquoten höher und damit Verzerrungen aufgrund von selektiver Teilnahme geringer sind. Dem kann jedoch entgegengehalten werden, dass Ergebnisdifferenzen auch Resultat anderer Effekte sein könnten, wie zum Beispiel der allgemeinen Rahmenbedingungen (Ausfüllen des Evaluationsbogens im Hörsaal unter Anwesenheit des Dozenten und der Kommilitonen versus Ausfüllen des Evaluationsbogens im häuslichen Umfeld). Auch zeigen eigene Ergebnisse, dass die Vorlesungsmittelwerte beider Erhebungsarten (mit Ausnahme der Bewertungen der Struktur der Vorlesungen) hochgradig korrelieren (siehe Tabelle 2). Sie zeigen darüber hinaus, dass Online-Evaluationen eine signifikant höhere Standardabweichung der Teilbereichsbewertungen aufweisen als Papier-Evaluationen. Dies könnte daran liegen, dass Online- und Papier-Evaluationen unter recht unterschiedlichen Rahmenbedingungen erfolgen. Aufgrund der Anwesenheit des Dozenten und der Kommilitonen besteht bei Papier-Evaluationen in gewissem Maße ein ‚Gruppenzwang‘ zur Teilnahme; die Teilnahme an der Online-Evaluation erfolgt dagegen völlig freiwillig in frei verwendbarer Zeit, die z. B. auch zum Lernen für die anstehenden Klausuren oder für Freizeitaktivitäten genutzt werden könnte. Von daher ist vorstellbar, dass Online-Evaluationen zu einer Selbstselektion der zur Teilnahme an der Evaluation berechtigten Studierenden führen, was bewirken könnte, dass

die Bewertungen differenzierter ausfallen. Für diese Erklärung spricht auch das Ergebnis von Layne/DeChristoforo/McGinty (1999, S. 229), dass die Teilnehmer der Online-Evaluation mehr Kommentare zu den Lehrveranstaltungen geschrieben haben. Die sich in den eigenen Daten andeutende Selbstselektion der zur Evaluation der jeweiligen Vorlesung berechtigten Studenten – sofern sich diese im Rahmen weiterer empirischer Untersuchungen mit größeren Fallzahlen verifizieren lässt – spricht dafür, dass Online-Befragungen die Ergebnisqualität von studentischen Evaluationen nicht zwangsläufig aufgrund geringerer Rücklaufquoten verschlechtern.

6.2 Evaluationszeitpunkt

Die Gesamtbewertungen von Vorlesungen zu verschiedenen Zeitpunkten sind – wie auch bei Kohlan – korreliert (siehe Tabelle 5; vgl. Kohlan 1973, S. 589). Bei Teilnehmerkonstanz und ausschließlicher Betrachtung der Gesamtbewertungen ergeben sich – wie auch bei Pruitt/Dicks/Tilley – keine signifikanten Unterschiede zwischen den Lehrbewertungen zu verschiedenen Zeitpunkten (siehe Tabelle 8; vgl. Pruitt/Dicks/Tilley 2011, S. 12). Die Teilbereichsbewertungen und die Ergebnisse bei variierender Teilnehmerzusammensetzung lassen aber durchaus einen Einfluss des Evaluationszeitpunkts erkennen. So sind bei Betrachtung der Vorlesungsmittelwerte die Gesamtbewertungen rund einen Monat vor der Klausur signifikant besser als in der letzten Vorlesungswoche oder nach der Klausur und im Falle signifikanter Abweichungen in den Bewertungen zu zwei Zeitpunkten wird im Durchschnitt stets der frühere Zeitpunkt besser bewertet. Dies stützt die bereits oben dargestellte Ansicht, dass bei vergleichender Analyse von Evaluationsergebnissen ein möglicher Einfluss des Evaluationszeitpunkts zu beachten ist (vgl. Reichmann 2007, S. 79). Die sich daraus ergebende Schlussfolgerung ist, dass studentische Lehrevaluationen zum gleichen Zeitpunkt erfolgen sollten, sofern die Ergebnisse miteinander verglichen werden sollen. Diese Empfehlung gilt nicht nur für Online-Evaluationen, sondern auch für Papier-Evaluationen. Zwar scheint es so, dass die in dieser Studie beobachteten Einflüsse des Evaluationszeitpunkts teilweise auf einem Selbstselektionseffekt beruhen. Allerdings zeigten sich auch bei Teilnehmerkonstanz Unterschiede (bezüglich der Teilbereichsbewertungen). Im Übrigen sind Selbstselektionseffekte auch bei Papier-Evaluationen nicht auszuschließen, auch wenn diese Online-Evaluationen vermutlich stärker betreffen. So ist vorstellbar, dass die Zusammensetzung der Lehrveranstaltungsteilnehmer und damit auch die Zusammensetzung der Befragungsteilnehmer bei der Papier-Evaluation systematisch zwischen verschiedenen Semesterzeitpunkten variiert. Denkbar ist beispielsweise, dass einige Studenten nur zum letzten Termin erscheinen, um etwaige Klausurhinweise nicht zu versäumen.

Ein Einfluss des Evaluationszeitpunkts auf die Differenziertheit der Bewertungen deutete sich hingegen nur für einen von sechs Vergleichen zweier Evaluationszeitpunkte an. Der Suche nach Indizien für einen derartigen Einfluss lag der Gedanke zugrunde, dass differenziertere Bewertungen einen höheren Zeitaufwand erfordern und dass Bewertungen von Vorlesungen somit je nach Evaluationszeitpunkt mehr oder weniger differenziert ausfallen könnten (z. B. kurz vor der Klausur weniger differenziert).

Ob sich der Evaluationszeitpunkt auf die Rücklaufquoten auswirkt, ist fraglich. Die Beobachtung, dass die geschätzten Rücklaufquoten bei Online-Evaluationen zu einem früheren Semesterzeitpunkt tendenziell (und für einige Zeitpunktvergleiche auch signifikant) höher sind als zu einem späteren Zeitpunkt, mag der mangelnden Bereitschaft von Studenten zur Mehrfachteilnahme geschuldet sein („Ermüdungseffekt“). Andererseits ist aber nicht auszuschließen, dass ein entsprechender Effekt auch unter „Realbedingungen“ (d. h. bei einmaliger Evaluation) zu beobachten wäre. Ein Grund dafür könnte sein, dass Studenten bei früherer Evaluation von etwaigen Lehrverbesserungen, die durch die Evaluationsergebnisse angestoßen werden, noch profitieren könnten, während die Veranstaltungskritik bei einem späteren Evaluationstermin vor allem (oder ausschließlich) nachfolgenden Studentengenerationen zugutekommt.

7. Fazit und Forschungsausblick

Die Ergebnisse dieser Studie lassen darauf schließen, dass Vergleiche von Evaluationen, die zu verschiedenen Zeitpunkten stattfanden, problematischer sind als Vergleiche von Evaluationen, bei denen verschiedene Erhebungsarten (Online- versus Papier-Fragebögen) eingesetzt wurden. Diese Beobachtung ist insofern von Interesse, als dass sich innerhalb einer Fakultät eher der Evaluationszeitpunkt als die Erhebungsart unterscheidet. Vor dem Hintergrund, dass studentische Lehrbewertungen zunehmend auch für Leistungsvergleiche der Lehrenden verwendet werden, ist von Bedeutung, dass ein früherer Evaluationszeitpunkt einen Wettbewerbsvorteil darzustellen scheint. Dies spricht dafür, dass studentische Lehrevaluationen zeitgleich (und wenn möglich auch auf die gleiche Art) erfolgen sollten, wenn Ergebnisvergleiche beabsichtigt sind.

In weiteren empirischen Untersuchungen sollte die Validität der Ergebnisse für größere Fallzahlen geprüft werden. Raum für weitere Forschung bietet auch die Frage, worauf der Einfluss des Evaluationszeitpunkts auf die Bewertungen von Vorlesungen zurückzuführen ist. Hier könnten Noteneffekte eine zentrale Rolle spielen. Im Hinblick auf den zunehmenden

Einsatz von Online-Evaluationen könnte in zukünftigen Studien nach weiteren Anhaltspunkten dafür gesucht werden, dass Teilnehmer an Online-Evaluationen differenziertere Bewertungen abgeben. Interessant wäre auch ein Vergleich von Online- und Papier-Evaluationen bei Teilnehmerkonstanz. Dabei würde sich allerdings das methodische Problem ergeben, dass Studenten möglicherweise Bedenken hinsichtlich der Wahrung der Anonymität bei der individuellen Zuordnung der Ergebnisse haben, was zu einer Verzerrung etwaiger Ergebnisunterschiede führen könnte.

Literatur

- Avery, Rosemary J./Bryant, W. Keith/Mathios, Alan/Kang, Hyojin/Bell, Duncan (2006): „Electronic course evaluations: does an online delivery system influence student evaluations?“, *Journal of Economic Education* 37 (1), S. 21-37.
- Brockx, Bert/Spooren, Pieter/Mortelmans, Dimitri (2011): „Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education“, *Educational Assessment, Evaluation and Accountability* 23 (4), S. 289-306.
- Brosius, Felix: „SPSS 19“, Heidelberg.
- Clayson, Dennis E./Haley, Debra A. (2011): „Are students telling us the truth? A critical look at the student evaluation of teaching“, *Marketing Education Review* 21 (2), S. 101-112.
- Felton, James/Koper, Peter T./Mitchell, John/Stinson, Michael (2008): „Attractiveness, easiness and other issues: student evaluations of professors on Ratemyprofessors.com“, *Assessment & Evaluation in Higher Education* 33 (1), S. 45-61.
- Kohlan, Richard G. (1973): „A comparison of faculty evaluations early and late in the course“, *Journal of Higher Education* 44 (8), S. 587-595.
- Layne, Benjamin H./DeCristoforo, Joseph R./McGinty, Dixie (1999): „Electronic versus traditional students rating of instruction“, *Research in Higher Education* 40 (2), S. 221-232.
- Meinefeld, Werner (2010): „Online-Befragungen im Kontext von Lehrevaluationen – praktisch und unzuverlässig“, *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62 (2), S. 297-315.
- o. V. (2012a): „Lehrbericht der Fakultät für Wirtschaftswissenschaften – Berichtszeitraum: Oktober 2010 bis September 2011“, Universität Regensburg, im Internet unter <http://www-wiwi.uni-regensburg.de/images/studium/Lb11.pdf> (letzter Aufruf am 15.07.2012).
- o. V. (2012b): „Online-Evaluation am Fachbereich Wirtschaftswissenschaften – Stand: Mai 2012“, Friedrich-Alexander Universität Erlangen-Nürnberg, im Internet unter http://www.wiso.uni-erlangen.de/studium/lehrevaluation/download/downloads/Online-Evaluation_Prozess-Timing-SS2012.pdf (letzter Aufruf am 15.07.2012).

- o. V. (2012c): „Der Ablauf des Verfahrens“, Wirtschaftswissenschaftliche Fakultät der Westfälischen Wilhelms-Universität Münster, im Internet unter <http://www.wiwi.uni-muenster.de/evaluation/lehrende/ablauf.html> (letzter Aufruf am 30.08.2012).
- o. V. (2012d): „Wie läuft die Evaluation der Veranstaltungen ab?“, Fakultät für Betriebswirtschaftslehre der Universität Mannheim, im Internet unter <http://www.bwl.uni-mannheim.de/de/qm/evaluationsstelle/faqs/#c284> (letzter Aufruf am 30.08.2012).
- o. V. (2012e): „Evaluation gestartet!“, Wirtschaftswissenschaftliche Fakultät der Julius-Maximilians-Universität Würzburg im Internet unter <http://www.economics.uni-wuerzburg.de/aktuelles/archiv/single/artikel/evaluation-11> (letzter Aufruf am 30.08.2012).
- Pruitt, J. Ross/Dicks, Michael R./Tilley, Daniel S. (2011): „Divergence and magnitude of student evaluations“, im Internet unter <http://ssrn.com/abstract=1887809> (letzter Aufruf am 02.08.2012).
- Reichmann, Gerhard (2007): „Überprüfung der Akzeptanz einer zentralen Evaluation von Lehrveranstaltungen am Beispiel der studentischen Lehrveranstaltungsbewertung an der Universität Graz“, Qualität in der Wissenschaft: Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration 1 (3), S. 77-81.
- Wettern, Michael (2008): „Lehrevaluation an Hochschulen“, Datenschutz und Datensicherheit 32 (1), S. 29-33.

Bisher erschienen:

Diskussionspapiere des Instituts für Organisationsökonomik

- DP-IO 8/2012** Effekte von Erhebungsart und -zeitpunkt auf studentische Evaluationsergebnisse
Laura Lütkenhöner
August 2012
- DP-IO 7/2012** Prolegomena zu einer Analyse ethischer und anderer Normen am Beispiel des Hochschulmanagements
Alexander Dilger
Juli 2012
- DP-IO 6/2012** The Impact of Physical Attractiveness on the Popularity of Female Tennis Players in Online Media
Stephanie Kiefer/Katrin Scharfenkamp
Juni 2012
- DP-IO 5/2012** Förderung von Wissenschaft zu nationalen und europäischen Fragen
Alexander Dilger
Mai 2012
- DP-IO 4/2012** Untersuchung von Indikatoren zur Qualitätsmessung von Reitschulen in Deutschland
Stephanie Kiefer
April 2012
- DP-IO 3/2012** Rigor, wissenschaftliche und praktische Relevanz
Alexander Dilger
März 2012
- DP-IO 2/2012** Socio-Demographic Characteristics and Human Capital of the German Federal Government's Members
Katrin Scharfenkamp/Alexander Dilger
Februar 2012
- DP-IO 1/2012** Die Zitationshäufigkeit als Qualitätsindikator im Rahmen der Forschungsleistungsmessung
Harry Müller
Januar 2012
- DP-IO 12/2011** Ein Forschungsleistungsranking auf der Grundlage von Google Scholar
Alexander Dilger/Harry Müller
Dezember 2011
- DP-IO 11/2011** Besonderheiten der Bewerbung um Promotionsstellen und -gelegenheiten
Alexander Dilger
November 2011
- DP-IO 10/2011** 1. Jahresbericht des Instituts für Organisationsökonomik
Alexander Dilger/Stephanie Kiefer/Katrin Scharfenkamp
Oktober 2011
- DP-IO 9/2011** Corporate Governance and Employee Power in the Boardroom
An Applied Game Theoretical Analysis
Benjamin Balsmeier/Andreas Bermig/Alexander Dilger/Hannah Geyer
September 2011

- DP-IO 8/2011** Ein Ranking von Hochschulen und (Bundes-)Ländern am Beispiel der Betriebswirtschaftslehre
Harry Müller/Alexander Dilger
August 2011
- DP-IO 7/2011** Befragung der Kommission Hochschulmanagement zu VHB-JOURQUAL
Alexander Dilger
Juli 2011
- DP-IO 6/2011** Director Interlocks and Executive Turnover in German Public Corporations
A Hazard Analysis for the Period from 1996 to 2008
Benjamin Balsmeier/Achim Buchwald/Alexander Dilger/Jörg Lingens
Juni 2011
- DP-IO 5/2011** Personalökonomik
Stärken, Schwächen und ihr Platz in der Personalwirtschaftslehre
Alexander Dilger
Mai 2011
- DP-IO 4/2011** Familienbewusste Personalpolitik und Unternehmenserfolg
Eine empirische Untersuchung
Christian Lehmann
April 2011
- DP-IO 3/2011** Welche Unternehmen berufen Vorstandsvorsitzende und andere Vorstände als externe Kontrolleure?
Eine empirische Analyse der Präsenz von externen Vorständen in den Aufsichtsräten deutscher Großunternehmen
Achim Buchwald
März 2011
- DP-IO 2/2011** Hat Julia aufgrund ihres Vornamens Wettbewerbsvorteile gegenüber Ayse und Chantal? Ein Experiment auf dem Beziehungs-, Nachhilfe- und Wohnungsmarkt
Laura Lütkenhöner
Februar 2011
- DP-IO 1/2011** Die dunkle Seite der Gerechtigkeit
Alexander Dilger
Januar 2011
- DP-IO 3/2010** On the Overconfidence-Effect in Teams
Hanke Wickhorst
Dezember 2010
- DP-IO 2/2010:** Leistung, Identifikation oder die Unsicherheit über den Spielausgang – was zählt wirklich? Relevante Einflussfaktoren auf die Zuschauerzahlen in der Basketball-Bundesliga
Hannah Geyer
November 2010
- DP-IO 1/2010:** A Citation Based Ranking of German-speaking Researchers in Business Administration with Data of Google Scholar
Alexander Dilger/Harry Müller
Oktober 2010



Herausgeber:
Prof. Dr. Alexander Dilger
Westfälische Wilhelms-Universität Münster
Institut für Organisationsökonomik
Scharnhorststr. 100
D-48151 Münster

Tel: +49-251/83-24303

Fax: +49-251/83-28429

www.wiwi.uni-muenster.de/io

