Diskussionspapierreihe
Working Paper Series

Helmut Schmidt
Universität
Universität der Bundeswehr Hamburg

MeinProf.de und die Qualität der Lehre

Klaus Beckmann
Andrea Schneider

Nr./ No. 90
Februar 2009

Department of Economics
Fächergruppe Volkswirtschaftslehre

Autoren / Authors

Klaus Beckmann
Helmut Schmidt Universität Hamburg / Helmut Schmidt University Hamburg
Institut für Finanzwissenschaft / Institute of Public Finance
Holstenhofweg 85
22043 Hamburg
Germany
klaus.beckmann@hsu-hh.de

Andrea Schneider
Helmut Schmidt Universität Hamburg / Helmut Schmidt University Hamburg
Institut für Finanzwissenschaft / Institute of Public Finance
Holstenhofweg 85
22043 Hamburg
Germany
andrea.schneider@hsu-hh.de

Redaktion / Editors

Helmut Schmidt Universität Hamburg / Helmut Schmidt University Hamburg
Fächergruppe Volkswirtschaftslehre / Department of Economics

Eine elektronische Version des Diskussionspapiers ist auf folgender Internetseite zu finden/
An electronic version of the paper may be downloaded from the homepage:
http://fgvwl.hsu-hh.de/wp-vwl

Koordinator / Coordinator

Kai Hielscher
wp-vwl@hsu-hh.de

Helmut Schmidt Universität Hamburg / Helmut Schmidt University Hamburg
Fächergruppe Volkswirtschaftslehre / Department of Economics

Diskussionspapier Nr. 90
Working Paper No. 90

MEINPROF.DE UND DIE QUALITÄT DER LEHRE

KLAUS BECKMANN
ANDREA SCHNEIDER

Zusammenfassung / Abstract

This paper compares student evaluation of teaching on the web site MeinProf.de to the standard evaluation procedures at German universities. While MeinProf offers the advantage of a broad, publicly available database, a number of theoretical and econometric problems emerge. Many of these problems are, however, pervasive in the student evaluation of teaching as a whole. We demonstrate that this seriously hampers empirical work on the relationship of teaching and research.

JEL-Klassifikation / JEL-Classification: I20, I23

Schlagworte / Keywords: Lehrevaluation, Lehre versus Forschung, Lehrqualität

Wir danken unseren Kollegen an der HSU, allen voran Annette Jünemann, Franco Reither und Klaus W. Zimmermann, für hilfreiche Kommentare zum Manuskript.

1. Einleitung

Es gibt gute Gründe, die Qualität der universitären Lehre zu messen: Ökonometrische Studien zum *Forschungs*output gibt es zuhauf (etwa Rauber und Ursprung 2008), nur fehlen demgegenüber quantitative Studien zur Lehre als dem zweiten großen Output von Hochschullehrern. Zwar besteht eine gewisse Evidenz, dass eine höhere Lehrbelastung (ein *Input*) *ceteris paribus* den Forschungsoutput reduziert (Fox und Milbourne 1999), doch liegen zu der spannenderen Frage einer Substitutionalität oder Komplementarität von Forschung und Lehre (als *Outputs*) kaum Studien vor.[1] Sind die produktiveren Forscher auch – bei gleichem Zeitaufwand – die besseren Lehrer, oder ist es umgekehrt? Für die empirische Überprüfung theoretischer Arbeiten, welche die Produktion in Hochschulen und die optimale Ausgestaltung von Dienstverhältnissen der Professoren unter Anwendung von Mehrprodukt-Prinzipal-Agent-Modellen (PA) untersuchen (Walckiers 2008), erscheint ein vernünftiges Maß für Qualität und Quantität des Lehroutputs nachgerade unverzichtbar.

Will man indes die Lehrqualität an den wirtschaftswissenschaftlichen Fakultäten in Deutschland genauer unter die Lupe nehmen, um z.B. eine Substitutionsbeziehung von Forschung und Lehre zu untersuchen oder theoretische Ergebnisse von PA-Modellen zur Vertragsbeziehung im Hochschulsektor empirisch zu verifizieren, stößt man zuallererst auf ein Verfügbarkeitsproblem:[2] Lehrevaluationen einzelner Veranstaltungen verbleiben zumeist – für die Allgemeinheit unveröffentlicht – in den Händen der Universitäten. Eine Abhilfe für dieses Problem scheint zunächst mit der Internetplattform MeinProf.de geschaffen.[3] Mit Hilfe dieser Internetplattform ist es für Studenten möglich, einzelne Lehrveranstaltungen in verschiedenen Kategorien nach dem Schulnotensystem zu beurteilen und eine Gesamtempfehlung für die jeweilige Veranstaltung abzugeben. Liegen genügend Bewertungen vor, so werden Durchschnittsnoten für Veranstaltungen und für die Lehrkräfte gebildet, die ihrerseits unter anderem für Rankings („Top 10“) genutzt werden. Die Internetpräsenz sieht sich hierbei immer wieder den Vorwürfen ausgeliefert, eine Plattform für rachsüchtige Studenten darzustellen und ein verzerrtes Bild der Lehrqualität abzubilden. Auch die Bedeutung der Dicke zu bohrender Bretter für die studentische Bewertung von Lehre wird MeinProf.de oft vorgeworfen (Freudenberger 2008).

In diesem Beitrag wollen wir beiden Behauptungen für den Bereich der Wirtschaftswissenschaften einmal genauer auf den Grund gehen. Zunächst bietet Kapitel 2 einen kleinen Überblick über die Bewertungen unter MeinProf.de und zeigt, dass die Hochschullehrer im Bereich der Wirtschaftswissenschaften keineswegs in einem schlechten Licht erscheinen. Kapitel 3 diskutiert dann die verschiedenen Vorwürfe, die Lehrevaluationen durch Studierende immer wieder gemacht werden. Es zeigt sich, dass einige dieser Vorwürfe sowohl MeinProf.de als auch die üblichen Befragungen im Hörsaal betreffen, während andere nur für

[1] Siehe allerdings Euwals und Ward (2005).

[2] Damit ist keineswegs gesagt, dass alle Schwierigkeiten überwunden wären, veröffentlichten die Universitäten nur ihre Evaluationsdaten. Im Abschnitt 3 werden wir uns noch kurz mit systematischen Problemen der Lehrevaluation befassen müssen.

[3] In den USA besteht ein recht ähnliches Angebot unter http://www.ratemyprofessors.com/. Vgl. Felton *et al.* (2008) für eine Analyse dieser Daten.

eine dieser Erhebungsformen gelten. MeinProf.de hat aus dieser Sicht auch Vorzüge zu bieten. Andererseits bestehen aber insgesamt erhebliche Zweifel, ob man die Qualität der Lehre überhaupt durch das Urteil der gegenwärtig Studierenden messen kann.

Anschließend vergleicht Kapitel 4 für die Helmut-Schmidt-Universität (UniBw) Hamburg und für die Leibniz-Universität Hannover im Bereich der Wirtschaftswissenschaften die MeinProf-Daten mit den Uni-intern durchgeführten Lehrevaluationen. Kapitel 5 fasst abschließend die Ergebnisse zusammen und beurteilt die Nutzbarkeit von MeinProf-Daten für ökonometrische Studien von Forschung und Lehre.

2. Die MeinProf-Daten im Bereich der Wirtschaftswissenschaften

Für unsere Analyse verfügen wir über kursbezogene Daten von Juli 2008 aus der MeinProf-Datenbank für alle Kurse, die insgesamt mindestens fünf Bewertungen erhalten haben.[4] Unser Datensatz umfasst 5413 Kurse von 3174 Lehrenden aus dem Bereich der Wirtschaftswissenschaften, wobei alle Hochschultypen vertreten sind, an denen wirtschaftswissenschaftliche Lehre angeboten wird. Für jede Lehrveranstaltungen liegen Teilnoten in sieben Kategorien – Fairness, Verständlichkeit, Material, Interesse, Spaß, Unterstützung und Note/Aufwand – nach dem Schulnotensystem vor. Aus den ersten sechs Kategorien, also ohne die Kategorie Note/Aufwand, wird dann eine Gesamtnote als ungewichteter Mittelwert errechnet.[5]

Die nachfolgende Tabelle 1 zeigt zunächst summarische Statistiken für die Ergebnisse in den sieben Kategorien und die Gesamtnote. Durchschnittlich 71 % der Respondenten empfahlen den jeweiligen Kurs.

Tab. 1: Bewertung wirtschaftswissenschaftlicher Kurse in MeinProf.de

	Fairness	**Verständl.**	**Material**	**Interesse**	**Spaß**	**Unterst.**	**Note/ Aufw.**	***Gesamt***
Mittelw.	2,15	2,29	2,37	2,50	2,60	2,28	2,49	2,37
Stdabw.	0,835	0,825	0,838	0,829	0,928	0,849	0,802	

Lehre wird in MeinProf.de im Durchschnitt folglich ganz ordentlich – „noch gut“ – bewertet, wobei das Urteil über den „Spaßfaktor“ am schlechtesten ausfällt und auch deutlich am stärksten streut. Betrachtet man die Verteilungen der Bewertungen im Detail, so fällt zunächst auf, dass diese keiner Normalverteilung folgen – in Abb. 1 anhand des Histogramms der (errechneten) Gesamtnoten exemplarisch dargestellt. Auch die Standardtechnik des Logarithmierens ergibt zwar *grosso modo* symmetrische Verteilungen, aber keine Normalität: Sowohl für die einzelnen Bewertungen als auch für deren Logarithmen führt ein Shapiro-Wilk-Test zur Ablehnung der Nullhypothese einer Normalverteilung.

[4] Die maximale Zahl der Bewertungen pro Kurs ist 82, der Durchschnitt 11,25. Die geringen Fallzahlen bei der Bewertung von Kursen stellen ein Problem von MeinProf.de im Vergleich zu der üblichen Methode der Lehrevaluation dar, Fragebögen im Kurs zu verteilen. Dadurch wird möglicherweise auch das Problem der Selbstselektion verschärft. Mehr dazu weiter unten in diesem Abschnitt.

[5] Zu Details siehe die Beschreibungen auf der Webseite http://www.meinprof.de/ selbst. Wir danken dem Team von MeinProf.de, allen voran Alexander Pannhorst, für die Überlassung der Daten.

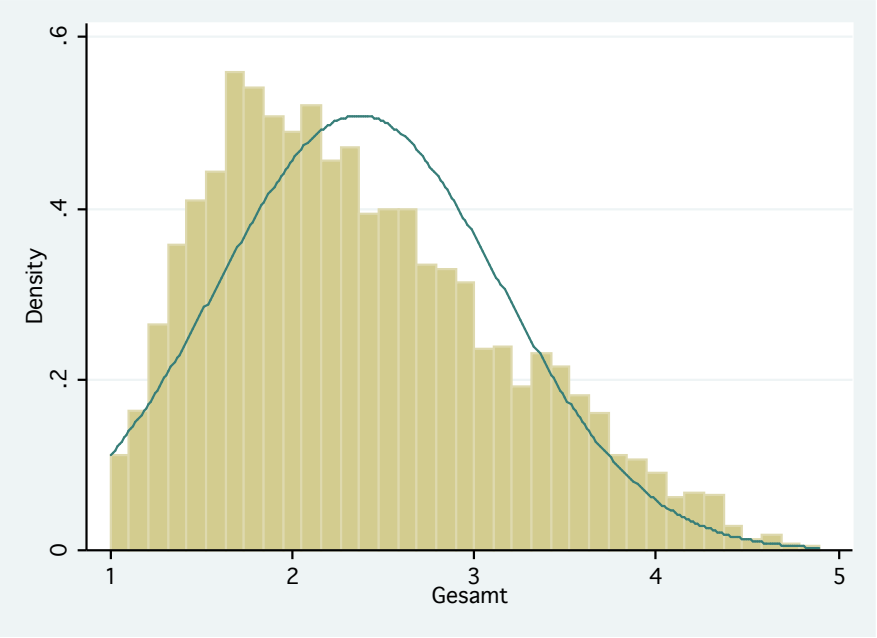

Abb. 1: Verteilung der Gesamtnoten bei MeinProf.de (Wirtschaftswissenschaften)

Dies mag auf den ersten Blick als ökonometrisch-technisches Problem erscheinen, doch steht dahinter eine noch fundamentalere Frage. Denn man verfügt über keinerlei Informationen über die Verteilungen der Bewertungen *in der Grundgesamtheit*, d.h. unter allen Studierenden. Wer aus dieser Gruppe eine Bewertung abgibt, bleibt der Spekulation überlassen (obgleich man es grundsätzlich durch eine gesonderte empirische Studie erfassen könnte). Im Unterschied dazu weiß man bei vielen anderen empirischen Problemen, welche Gruppe im Sample über- bzw. unterrepräsentiert ist.

Ein solcher *Sample selection bias* plagt die typischen Lehrevaluationen, die auf der Erfassung von im Hörsaal ausgefüllten Fragebögen basieren, weit weniger. Dennoch werden diese Probleme auch hier nicht völlig ausgeräumt: Beispielsweise können Studierende auf das Abgeben des Bogens verzichten, und weil man erst am Ende eines Semesters befragen kann – so dass die Studierenden die Veranstaltung auch beurteilen können –, hat man die völlig Frustrierten möglicherweise bereits verloren. Man kann erwarten, dass die Bewertungen in der üblichen Lehrevaluation aufgrund dieser Effekte nach oben verzerrt sind. Für MeinProf.de lässt sich weit weniger sagen.

Ein zweites, nicht ganz so bedeutsames Problem kann durch die Heterogenität der Hochschulen in der Studie entstehen – beispielsweise legen einige mehr Wert auf Forschung, andere auf Lehre.[6] Stützen sich Studierende bei der Bewertung von Lehre insbesondere auf die *relative* Leistung im Vergleich mit anderen Professoren, die sie erlebt haben, kann dies den Vergleich zwischen den Evaluationsergebnissen an verschiedenen Hochschulen bzw. deren Nutzung in einem Modell, das auch den Forschungsoutput berücksichtigt, verzerren.

[6] Diese Differenzierung dürfte im angloamerikanischen Hochschulsektor weit stärker ausgeprägt sein als im deutschsprachigen Raum.

Dieser Einwand träfe die herkömmlichen Lehrevaluationen ebenso wie MeinProf.de.

Allerdings finden wir in unserem Datensatz Anzeichen für systematische Unterschiede in den Urteilen nach Hochschultypen.[7] Bilden wir eine Dummyvariable die den Wert Eins für eine Bewertung mit besser als 2,0 und sonst den Wert Null annimmt, dann erweist sich, dass die Lehre an Fachhochschulen und Akademien mit signifikant höherer Wahrscheinlichkeit gut bewertet wird als an wissenschaftlichen Hochschulen (siehe Tabelle 2, Pearson'sches χ^2 = 39,96, signifikant auf dem 1 %-Niveau). Dies dürfte nicht der Fall sein, wenn Studierende bei MeinProf.de ausschließlich nach *relativen* Maßstäben urteilten.

Tab. 2: Anteile „guter" Gesamtbewertungen nach Hochschultypen

	Akademie	FH	Uni
Gesamtnote bis 2,0	60 %	58,8 %	67,8 %
Gesamtnote besser als 2,0	40 %	41,2 %	32,2 %

Ein drittes Problem betrifft die Korrelation der Bewertungsvariablen bei MeinProf.de untereinander. Wie die Korrelationsmatrix (Tabelle 3 unten) zeigt, sind die Bewertungen in den verschiedenen Kategorien in der Tat hochgradig korreliert – im Übrigen noch etwas stärker, als es Felton et al. (2008) für die Daten von ratemyprofessor.com finden. Die erste Konsequenz betrifft ökonometrische Schätzungen mit diesen Daten: Weil der Einfluss der verschiedenen Variablen schlecht separiert werden kann, ist es besser, nur eine davon – etwa die errechnete Gesamtnote – in das Modell aufzunehmen. Ansonsten besteht nicht notwendig ein gravierendes Problem (Kennedy 2008).

Der wesentliche Punkt hier ist aber weniger ein methodischer als ein inhaltlicher: Es wird gerne behauptet, dass eine eindeutige *Kausalitäts*beziehung zwischen den Teilnoten bestehe, weil Studierende ihre Gesamteinschätzung eines Kurses (die sich in den verschiedenen Teilnoten widerspiegele) allein von dem *eigenen erwarteten Abschneiden in Relation zum geleisteten Arbeitseinsatz* abhängig machten. Jüngst kritisiert Freudenberger (2008)[8] MeinProf.de mit dem Hinweis, dass die Studierenden bei der Bewertung ihrer Hochschullehrer die MeinProf-Variable „Verhältnis von Aufwand und Note" in den Mittelpunkt rücken, und dass dies wohl kaum etwas mit der Qualität ihres Lehrers zu tun haben könne. Obwohl er eine Multikollinearität kurz als weiteren Erklärungsansatz anspricht, erstellt Freudenberger (2008) leider keine Korrelationsmatrix (siehe Tabelle 3). Diese zeigt jedoch, dass *alle* sieben Teilnoten bei MeinProf.de stark korrelieren. Und der Vergleich der kursiv gesetzten Werte mit den übrigen Inhalten einer Spalte verdeutlicht, dass die Variable Note/ Aufwand beim paarweisen Vergleich nirgendwo die höchste Korrelation aufweist, also mit *keiner* der anderen Variablen am stärksten korreliert.

[7] Es ist ein Vorzug der hochschulübergreifenden Erhebung nach einem einheitlichen System, dass wir solche Vergleiche überhaupt anstellen können. Nicht-standardisierte Evaluationsdaten, selbst wenn sie verfügbar wären, gestatteten dies nicht. Ein Punkt für MeinProf.de.

[8] Grundlage bilden hier allerdings nur aus dem Internet ausgelesene Daten für einen Teil des insgesamt verfügbaren Samples, nämlich die Bewertungen von 99 Veranstaltungen am Fachbereich Wirtschaftswissenschaften der FH Mainz aus dem April 2008.

Tab. 3: Korrelationsmatrix der Teilnoten bei MeinProf.de (Wirtschaftswissenschaften)

	Ges	**Fair**	**Verstdl**	**Mat**	**Interesse**	**Spaß**	**Unterst**	**Aufw**
Gesamt	1							
Fairness	0,893	1						
Verständlich	0,932	0,764	1					
Material	0,886	0,739	0,818	1				
Interessant	0,950	0,801	0,880	0,795	1			
Spaß	0,920	0,758	0,853	0,727	0,910	1		
Unterstützung	0,948	0,882	0,842	0,831	0,866	0,822	1	
Note / Aufwand	0,809	*0,838*	*0,734*	*0,662*	*0,739*	*0,725*	0,775	1

Während diese Beobachtung keinesfalls ein sicheres Urteil gestattet, so sind unsere Daten jedenfalls auch mit anderen Hypothesen kompatibel. So etwa mit der alternativen Vermutung, dass Studenten, wenn sie eine Veranstaltung insgesamt gut finden, auch schlechte Materialien etc. nicht ganz so hart bewerten. Nach dieser Lesart gehen Studenten – umgekehrt zum MeinProf.de Bewertungssystem – zunächst von ihrem Gesamteindruck aus und vergeben dann unter diesem Einfluss die Noten in den einzelnen Kategorien.

Der oben zitierte Vorwurf, Studierende richteten sich bei der Evaluation der Lehre (zu) stark nach der Dicke des zu bohrenden Brettes, ist freilich nicht MeinProf-spezifisch, sondern wird seit langem gegenüber der Student Evaluation of Teaching insgesamt erhoben. Wir können uns also die Frage stellen, ob der Gedanke überhaupt taugt, die Qualität der Lehre durch Befragung von Studierenden erheben zu wollen. Beginnen wir dazu mit einigen Hinweisen zu der einschlägigen Debatte!

3. Lehrqualität und ihre Messung

Wie kann man Qualität von Lehre operationalisieren? Während in vielen hochschulpolitisch orientierten Schriften einer präzisen Antwort auf diese Frage eher ausgewichen wird – beispielsweise lässt es der Wissenschaftsrat (2008, 17-19) bei einem Hinweis auf die unterschiedlichen Standpunkte der Stakeholder bewenden, ohne sich festzulegen –, geht das für unsere Zwecke nicht an. Drei Ansätze der Operationalisierung liegen besonders nahe:

- durch den beruflichen Erfolg der Absolventen (Lüdeke und Beckmann 2001) – und die Erfüllung sonstiger Bildungsziele.[9] So oder so gelangt man so zu einer *Befragung der Alumni*.
- durch das Abschneiden der Studierenden bei standardisierten Tests (etwa PISA, TAMSS) oder Zentralprüfungen (Bayerisches Abitur, juristische Staatsprüfungen). Der Vorteil ist eine standardisierte Erfassung des Outputs, der Nachteil, dass hier auch

[9] Berücksichtigt man dabei neben dem Mittelwert auch noch die Streuung? Für die PISA-Daten wird das oft gemacht mit dem impliziten Urteil, dass eine geringere Streuung bei gleichem Mittelwert vorzuziehen sei. Wenn aber der gesellschaftliche Fortschritt vor allem durch Extremleistungen getrieben wird (Sinn 2008), kann das Urteil auch anders ausfallen.

akademische Leistungen als Indikatoren für die letztlich zu messende Vermittlung von Fähigkeiten genommen werden (also den Erfolg der Alumni).

- durch die Befragung von Studierenden.

Von den genannten Möglichkeiten ist sicherlich die dritte die schwächste.

Die Debatte über die Verlässlichkeit der *Student Evaluation of Teaching* (SET) tobt schon seit Jahrzehnten und hat in den USA, wo SET eine nicht unwesentliche Rolle bei der Einräumung von *Tenure* oder den Leistungszulagen der Professoren spielt, eine erhebliche hochschulpolitische Brisanz. Im Mittelpunkt steht meist der Vorwurf, dass die SET Anreize für die Lehrkräfte schafft, Lehrinhalte und Prüfungsanforderungen zu verändern. Insbesondere macht man die Lehrevaluation in den Vereinigten Staaten für eine schleichende Verbesserung des Notenniveaus (die „*grade inflation*“) verantwortlich: Studierende bewerten Professoren *ceteris paribus* um so besser, je besser ihre eigene erwartete Abschlussnote ist (Greenwald und Gillmore 1997). Sabot und Wakeman-Linn (1991) zeigen empirisch, dass Studenten umgekehrt auch bei der Wahl der Studiengänge und Kurse das durchschnittliche Notenniveau einfließen lassen. Wo die Auslastung von Kursen oder deren Bewertung in der studentischen Evaluation ein Kriterium für die Mittelvergabe ist, haben Professoren unter sonst gleichen Umständen einen stärkeren Anreiz, Studenten die Prüfung aktiv zu erleichtern.[10]

Jüngst haben Weinberg, Fleisher und Hashimoto (2007) das Abschneiden von Studierenden in späteren Studienabschnitten als Indikator für den Lernerfolg in Einführungskursen in Abhängigkeit von verschiedenen Charakteristika der Einführungskurse, darunter der individuellen Lehrevaluation und der tatsächlichen Note, geschätzt. Dabei zeigte sich eine Korrelation zwischen Kursnote und Lehrevaluation, doch erwies sich der Parameter der Evaluationsvariable als insignifikant bei der Schätzung des Lernerfolges, sobald für die Kursnote kontrolliert wurde.

Andere Studien legen die Existenz weiterer Determinanten für Evaluationsergebnisse nahe, die sich nicht ohne weiteres als relevante Dimensionen von Lehrqualität deuten lassen – insbesondere die „Coolness“ des Professors. Williams und Ceci (1997) untersuchen ein Realexperiment, in dem sich zwei Kurse nur durch den Enthusiasmus der Präsentation unterschieden, und fanden signifikant bessere Bewertungen in dem rhetorisch überzeugenderen Kurs. Daneben finden sie auch Indizien für eine positive Korrelation zwischen den Prüfungsergebnissen der Studenten und ihrer Bewertung der Lehrkraft. In einem jüngeren Papier verwenden Felton *et al.* (2008) die Daten des amerikanischen Vorbilds für MeinProf.de, http://www.ratemyprofessor.com/, und zeigen, dass insbesondere eine „Hotness“-Variable stark mit dem Gesamtergebnis korreliert und andere Variablen, darunter auch Qualität und Leichtigkeit des Erfolgs, für „attraktivere“ Professoren signifikant höher ausfallen.[11]

[10] Dies ist nicht der einzige Anreiz. Immerhin ist ein Student mit bestandener Prüfung kein zweites Mal zu prüfen, und Studierende mit besseren Noten dürften unter sonst gleichen Umständen auch sonst weniger Arbeit machen. Dem steht aus Sicht eines an der Verringerung des Arbeitsleids interessierten Professors entgegen, dass Kurse mit einem „leichten“ Ruf mehr Studenten anziehen (Sabot und Wakman-Linn 1991).

[11] „Hotness“ können Studierende bei RateMyProfesor.com auf einer dreistufigen (Likert-)Skala ausdrücken: negativ, neutral und positiv. Ein direktes Pendant dafür gibt es bei MeinProf.de nicht – am nächsten erscheint uns die „Spaß“-Variable (als Schulnote) und die Gesamtempfehlung (ein Dummy).

Schließlich zeigt Sproule (2002), dass das ökonometrische Modell des Lernerfolgs, welches studentischen Lehrevaluationen zugrunde liegt, fehlspezifiziert ist, weil ihm bestimmte unabhängige Variablen unter Kontrolle der Auszubildenden (etwa deren Anstrengungen) und der Administration (etwa die Ausstattung) fehlen.[12]

Stellt man MeinProf.de den üblichen Lehrevaluationen gegenüber, so zeigt sich, dass beide eine Reihe von Schwächen – vor allem die von Greenwald und Gillmore (1997) hervorgehobene Abhängigkeit vom erwarteten Erfolg in der Klausur und dem „Vorkauen" von Inhalten – teilen. MeinProf.de hat gegenüber den üblichen Lehrevaluationen durch Befragen in der Vorlesung den Vorteil einer bundesweit einheitlichen Erfassungsmethodik und standardisierter Befragungsbedingungen, dem allerdings aus theoretischer Perspektive der erhebliche Nachteil einer möglichen Selbstselektion der Probanden[13] gegenübersteht: Man könnte vermuten, dass sich besonders diejenigen Studierenden die Zeit zu einer Bewertung nehmen, die vom Professor begeistert sind oder aber die Veranstaltung entnervend schlecht finden. Solchen Vermutungen gilt es nun nachzugehen.

4. Hält MeinProf.de dem Vergleich mit üblichen Lehrevaluationen stand?

Im folgenden vergleichen wir MeinProf-Daten für die Kurse zweier wirtschaftswissenschaftlicher Fakultäten mit deren Lehrevaluation. Diese Beschränkung ist der Datenverfügbarkeit geschuldet: Im Falle Hannover sind die Ergebnisse im Internet veröffentlicht. Bei der HSU, unserer Heimatuniversität, lagen uns die Rohdaten vor. Selbstverständlich wurden die Datensätze vor der Auswertung anonymisiert.

Die Leibniz-Universität Hannover

Auf Grundlage der Lehrevaluationen der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover für die Sommersemester 2004, 2006 und 2007 sowie für die Wintersemester 2004/2005 und 2005/2006 lässt sich ein erster Vergleich von Lehrevaluationsdaten und MeinProf-Daten erstellen. Um möglichst viele Daten nutzen zu können, dient als Vergleichsgegenstand die Gesamtbewertung eines einzelnen Kurses (über verschiedene Semester hinweg), den eine Lehrperson abgehalten hat, und nicht die Gesamtbewertung der Lehrperson an sich.

Für die damit zur Verfügung stehenden 21 Beobachtungen (Kurse) ergibt sich eine Kreuzkorrelation der Gesamtnote der Lehrevaluationen und der MeinProf-Gesamtnote von 0,6556.; diese Korrelation ist überdies auf dem 1 %-Niveau signifikant. Augenscheinlich besteht durchaus ein, wenn auch kein perfekter, linearer Zusammenhang zwischen Gesamtnote der Lehrevaluation und Gesamtnote der MeinProf-Daten. Es zeigt sich allerdings, dass in den vorderen Teil der Notenskala (Gesamtnote zwischen 1.0 - 1.7) die MeinProf-Bewertung

[12] Die fehlende Definition von Qualität in der Lehre, Unklarheit über die zu verfolgenden Ziele und methodische Probleme der verwendeten Evaluationsverfahren werden allerdings auch bei alternativen Verfahren wie Lehrberichten und peer reviews der Lehre bemängelt (Kieser et al. 1996). Vgl. auch Emery, Kramer und Tian (2001).

[13] Freilich können sich auch Professoren selbst selektieren, weil für Hochschullehrer die Möglichkeit besteht, ihren Eintrag bei MeinProf.de streichen zu lassen. Das legt die Vermutung nahe, dass besonders schlecht abschneidende Professoren nicht mehr im Sample enthalten sind.

tendenziell besser ausfällt als die Bewertung mittels der Lehrevaluationen. Im hinteren Bereich der Notenskala (Gesamtnote 3.0 – 4.0) fallen die MeinProf-Gesamtnoten tendenziell schlechter aus als die Bewertungen der Lehrevaluationen. Die MeinProf-Bewertungen zeigen somit eine stärkere Gewichtung der Extreme.

Versucht man die Gesamtnote der Lehrevaluation (gesno_{le}) durch die Gesamtnote der MeinProf-Daten (gesno_{mp}) mittels linearer Regression zu erklären, ergibt sich die folgende Regressionsgerade

$$\begin{array}{lll} \text{gesno}_{\text{le}} & = 0,312\ \text{gesno}_{\text{mp}} & +1,439 \\ & \quad (3,78) & \quad (6,51) \end{array} \tag{1}$$

mit den t-Werten in Klammern.

Auch wenn sich die durchschnittlichen Gesamtnoten beider Erhebungen mit 2,48 (MeinProf.de) und 2,23 (Lehrevaluationen) nur geringfügig unterscheiden, so sind die Bewertungen von MeinProf.de jedoch stärker gestreut als die Bewertungen der Lehrevaluationen. Eine Ursache hierfür dürfte die geringe Anzahl an Bewertungen (im Durchschnitt 9,66 pro Kurs) der MeinProf-Erhebungen sein. Für die uni-intern durchgeführten Evaluationen ergibt sich hier immerhin eine durchschnittliche Bewertungsanzahl von 138.

Zahlreiche einschlägige Studien, etwa die Analyse von RateMyProfessor.com-Daten bei Felton et al. (2008) oder auch Freudenberger (2008), heben auf Korrelationen zwischen Bewertungen ab. Bei näherem Hinsehen kann dies aber nicht genügen, und zwar auch dann nicht, wenn man zwar nur bivariate Zusammenhänge betrachtet, aber Evaluationsdaten von mehreren Hochschulen zusammen fasst.[14] Denn dabei können gleich eine Reihe von Problemen auftreten:

- Zunächst könnten die Daten schlicht *unkorreliert* sein, so dass kein signifikanter linearer Zusammenhang zwischen den beiden Bewertungen bestünde.
- Sind die Daten korreliert, kann der lineare Zusammenhang – die geschätzte Regressionsgerade – noch von der Winkelhalbierenden abweichen (siehe die Abbildung 2 unten). Das mag an verschiedenen Bewertungsskalen, aber auch an anderen Eigenschaften der Evaluationsverfahren liegen. Zunächst handelt es sich hier nicht um ein Problem, ist eine positiv affine Transformation doch für lineare Regressionsanalysen unbedenklich, allerdings müsste zur Bildung eines aggregierten Datensatzes im Idealfall der lineare Zusammenhang *für alle Hochschulen gleich* sein. Davon aber ist schon deshalb nicht auszugehen, weil neben der bekannten Schulnotenskala auch andere Skalen gebräuchlich sind.
- Schließlich könnten drittens Ausreißer die Messung erschweren, was insbesondere angesichts der geringen Fallzahlen pro Kurs in den MeinProf-Erhebungen problematisch erscheint.

[14] Weitere Herausforderungen stellen sich ein, wenn man solche Daten für multivariate Schätzungen nutzt. Für unsere Betrachtung kann dies dahinstehen.

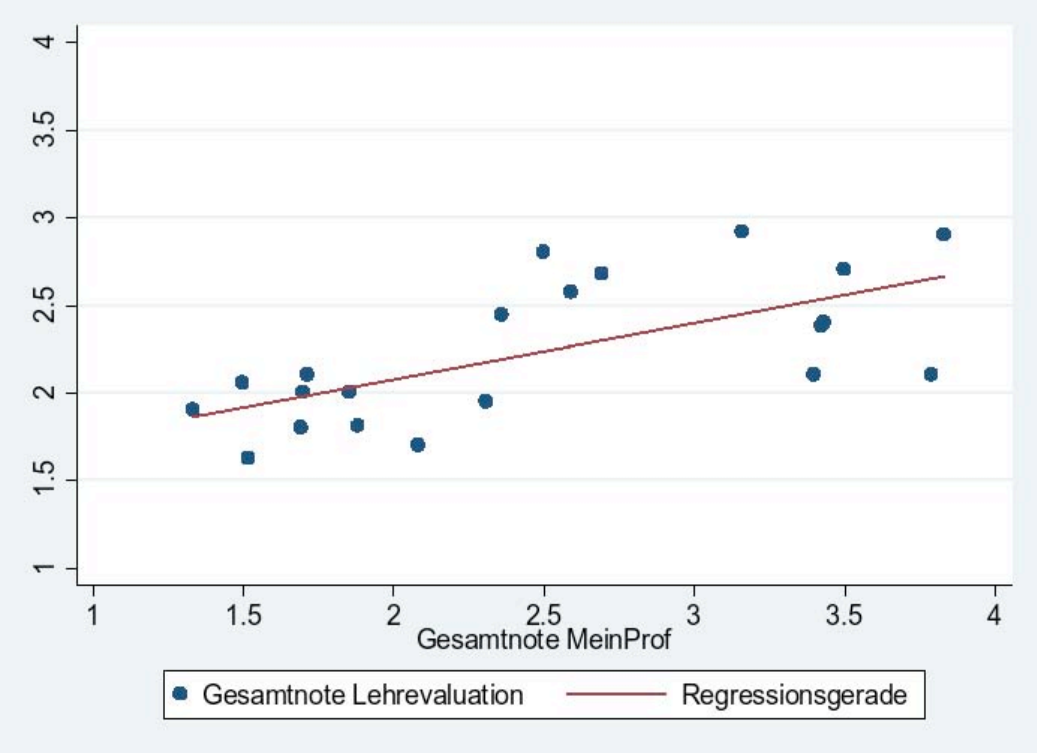

Abb. 2: Scatterplot der Gesamtnoten der Lehrevaluationen und von MeinProf.de für die Wirtschaftswissenschaftliche Fakultät Hannover

Vorerst bleibt festzuhalten, dass selbst beim Vorliegen eines linearen Zusammenhangs von MeinProf-Daten und den Daten der Lehrevaluationen die MeinProf-Daten nicht ohne weiteres als Proxy für die Lehrqualität in einer Panel-Studie mit verschiedenen Lehreinrichtungen genutzt werden dürfen. Dies wäre nur dann zulässig, wenn für alle Lehreinrichtungen *dieselbe* lineare Transformation zugrunde liegen würde.

Die HSU

Für die Helmut-Schmidt-Universität (UniBw) Hamburg lagen uns sämtliche an der Fakultät für Wirtschafts- und Sozialwissenschaften im Studienjahr 2007/8 durchgeführten Lehrevaluationen vor. Von den insgesamt 33 Beobachtungen verfügen wir für 19 Kurse auch über MeinProf-Daten mit durchschnittlich 6,42 Bewertungen pro Kurs, die wir im folgenden nutzen.[15] Dabei werden die Leistungen nicht wie zuvor durch Schulnoten, sondern auf einer Skala von 1 (schlecht) bis 7 (gut) bewertet, so dass die Regressionsgerade bei dem erwarteten Zusammenhang fällt. Die Korrelation zwischen den MeinProf- und den Evaluationsdaten ist recht hoch, und auch der lineare Zusammenhang erscheint zufrieden stellend (siehe Abb. 3 unten). Allerdings wird die Regressionsgerade durch zwei Ausreißer „gedreht“ – ohne diese lägen sämtliche verbleibenden Datenpunkte im 95 %-Konfidenzintervall.

[15] An der Fakultät für Wirtschafts- und Sozialwissenschaften der HSU HH bestehen insgesamt 38 Professuren, und es werden je drei Bachelor- und Master-Studiengänge (BWL, VWL, Politikwissenschaften) angeboten. In der Stichprobe sind Kurse aus den Wirtschaftswissenschaften, der Politikwissenschaft, den Rechtswissenschaften, der Mathematik und den Verwaltungswissenschaften enthalten.

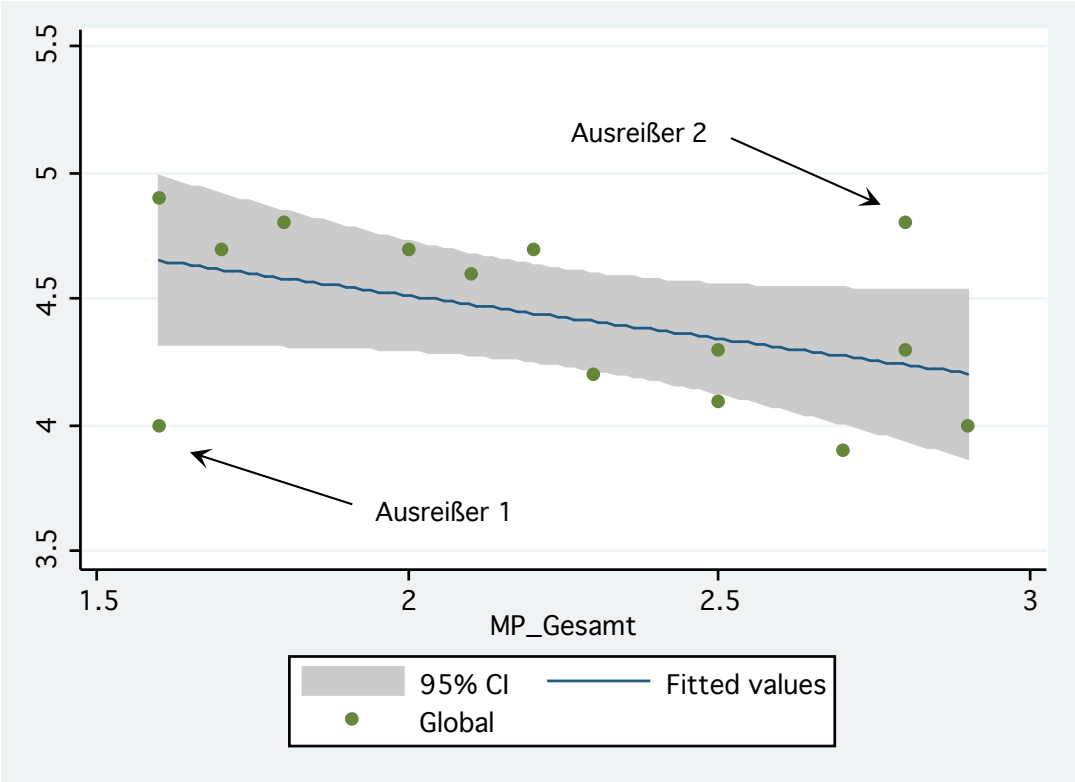

Abb. 3: MeinProf-Gesamtnote versus Gesamturteil in der Lehrevaluation (Fak. WISO an der HSU HH, Studienjahr 2007/8)

Bei beiden Ausreißern handelt es sich um kleine Lehrveranstaltungen mit der Mindestzahl erfasster Bewertungen in MeinProf.de, bei denen die Leistung der Dozenten im einen Falle durch die MeinProf-Mitglieder wesentlich besser, in dem anderen Fall wesentlich schlechter eingestuft wird als durch die Lehrevaluation in der Veranstaltung. Dies sind möglicherweise Fälle, in denen sich bei MeinProf.de gerade stark emovierte Personen zu Wort gemeldet haben.

Einen ökonometrischen Beleg für derlei Spekulation geben die Daten freilich nicht her, und es ist auch nicht absehbar, wie sich mit nur kleinen Änderungen am Design die Grundlage für ein Urteil schaffen ließe. Für Hannover können wir zwar keine solche Ausreißer identifizieren, haben andererseits keine Gewissheit, dass die genannten Effekte nicht auftreten. Hier besteht in der Tat ein großes Problem in der Vergleichbarkeit von MeinProf-Daten und Lehrevaluationsdaten der einzelnen Universitäten.

5. Fazit

Lässt man die kurzen Ausführungen Revue passieren, muss man zunächst wohl doch festhalten, dass es keinen Grund gibt, MeinProf.de im Vergleich zur anderen Lehrevaluatio-

nen zu verteufeln. Wir finden hier kein Tollhaus rachsüchtiger Durchfaller einerseits und fauler Fanbuben andererseits, welche in der Mehrzahl überzogene Ansichten „ins Internet stellen“ und extreme Bewertungen registrieren. Wo wir die Evaluationen direkt vergleichen können, sind die Bewertungen einigermaßen korreliert und liegen bis auf wenige Ausreißer innerhalb der 95%-Konfidenzintervalle der jeweils anderen Quelle.

Doch das genügt nicht. Wir mögen keine praktischen Anhaltspunkte für das oft vermutete Selbstselektionsproblem in den MeinProf-Daten finden können, aber die Möglichkeit bleibt aufgrund unserer geringen Basis für Vergleiche bestehen. Und darüber hinaus haben alle Formen der SET (Lehrevaluation durch Studierende) Probleme gemeinsam, für die wir auch Anzeichen finden, ganz abgesehen von der erheblichen Multikollinearität der Bewertungsvariablen. Gerade in dieser Hinsicht weisen die Daten – auch wenn wir die behauptete Dominanz der Kategorie „Note/Aufwand“ nicht bestätigen können – gravierende Mängel auf.

Für den empirischen Bildungsökonomen, der seine bibliometrischen Forschungsdaten dringend um eine metrisch oder rangskalierte Variable der Lehrqualität ergänzen muss, bietet MeinProf.de keine Lösung. Die üblichen Lehrevaluationen freilich auch nicht – selbst wenn man die persistierende Geheimhaltung überwinden könnte, verblieben noch genügend Herausforderungen, um diese Daten nicht ohne weiteres verwenden zu können. Die Lehre scheint sich der Bildungsökonomik auf der Leistungsseite zu entziehen.

Literatur

Emery, Charles, Tracy Kramer und Robert Tian (2001). Customers vs. products: adopting an effective approach to business students. *Quality Assurance in Education* 9, 110–15.

Euwals, Rob und Melanie E. Ward (2005). What matters most: teaching or research? Empirical evidence on the remuneration of British academics. *Applied Economics* 32, 1655-1672.

Felton, James, Koper, Peter T., Mitchell, John B. and Stinson, Michael (2008). Attractiveness, Easiness and Other Issues: Student Evaluations of Professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 33 (1), 45-61.

Fox, Kevin J. und Ross Milbourne (1999). What determines research output of academic economists? *The Economic Record* 75 (230), 256-267.

Freudenberger, Axel (2008). Motive bei studentischen Evaluationen von Lehrleistungen unter „MeinProf.de“ – Ein auffälliger Zusammenhang in den Daten der Website. *WiSt Wissenschaftliches Studium*, 11, 617-619.

Greenwald, Anthony G. und Gerald M. Gillmore (1997). No Pain, No Gain? the Importance of Measuring Course Workload in Student Ratings of Instructions. *Journal of Educational Psychology* 89: 743–51.

Kennedy, Peter (2008). *A Guide to Econometrics*. 6. Aufl. MIT Press.

Kieser, A., E. Frese, D. Müller-Böling und N. Thom (1996). Probleme der externen Evalua-

tion wirtschaftswissenschaftlicher Studiengänge. *Zeitschrift für Betriebswirtschaft* 1/96, 69–93.

Lüdeke, Reinar und Klaus Beckmann (2001). Die Passauer Absolventenstudie "Wirtschaftswissenschaften": Leistungsindikatoren (Noten), Einkommensniveaus, Einkommensprofile und Einkommensbarwerte, in: R.K. v. Weizsäcker (Hg.), Bildung und Beschäftigung (Schriften des Vereins für Socialpolitik). Berlin: Duncker&Humblot, 27–122.

Rauber, Michael und Heinrich W. Ursprung (2008). Life Cycle and Cohort Productivity in Economic Research: The Case of Germany. *German Economic Review* 9 (4), 431-456.

Sabot, Richard und John Wakeman-Linn (1991). Grade Inflation and Course Choice. *Journal of Economic Perspectives* 5, 159–70.

Sinn, Hans-Werner (2008). Barbaren oder Gelehrte vor den Toren? *ifo-Standpunkte* 97/2008.

Sproule, Robert A. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review* 21, 287–94.

Walckiers, Alexis (2008). Multi-dimensional contracts with task-specific productivity: an application to universities. *International Tax and Public Finance*, 15, 165-198.

Weinberg, Bruce A., Belton M. Fleisher und Masanori Hashimoto (2007). Evaluating methods for evaluating instruction: the case of higher education. *NBER Working Paper* W12844.

Williams, Wendy M. und Stephen J. Ceci (1997). How'm I Doing? *Change* Sep/Oct (1997): 13–23.

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

DISKUSSIONSPAPIERE DER FÄCHERGRUPPE VOLKSWIRTSCHAFTSLEHRE

DISCUSSION PAPERS IN ECONOMICS

Die komplette Liste der Diskussionspapiere ist auf der Internetseite veröffentlicht / for full list of papers see: http://fgvwl.hsu-hh.de/wp-vwl

2009

90 Beckmann, Klaus; Schneider, Andrea. MeinProf.de und die Qualität der Lehre, Februar 2009.

89 Berlemann, Michael; Hielscher, Kai. Measuring Effective Monetary Policy Conservatism, February 2009.

88 Horgos, Daniel. The Elasticity of Substitution and the Sector Bias of International Outsourcing: Solving the Puzzle, February 2009.

87 Rundshagen, Bianca; Zimmermann, Klaus W.. Buchanan-Kooperation und Internationale Öffentliche Güter, Januar 2009.

2008

86 Thomas, Tobias. Questionable Luxury Taxes: Results from a Mating Game, September 2008.

85 Dluhosch, Barbara; Zimmermann, Klaus W.. Adolph Wagner und sein „Gesetz“: einige späte Anmerkungen, August 2008.

84 Zimmermann, Klaus W.; Horgos, Daniel. Interest groups and economic performance: some new evidence, August 2008.

83 Beckmann, Klaus; Gerrits, Carsten. Armutsbekämpfung durch Reduktion von Korruption: eine Rolle für Unternehmen?, Juli 2008.

82 Beckmann, Klaus; Engelmann, Dennis. Steuerwettbewerb und Finanzverfassung, Juli 2008.

81 Thomas, Tobias. Fragwürdige Luxussteuern: Statusstreben und demonstratives Konsumverhalten in der Geschichte ökonomischen Denkens, Mai 2008.

80 Kruse, Jörn. Hochschulen und langfristige Politik. Ein ordnungspolitischer Essay zu zwei Reformutopien, Mai 2008.

79 Kruse, Jörn. Mobile Termination Carrier Selection, April 2008.

78 Dewenter, Ralf; Haucap, Justus. Wettbewerb als Aufgabe und Problem auf Medienmärkten: Fallstudien aus Sicht der „Theorie zweiseitiger Märkte“, April 2008.

77 Kruse, Jörn. Parteien-Monopol und Dezentralisierung des demokratischen Staates, März 2008.

76 Beckmann, Klaus; Gattke, Susan. Status preferences and optimal corrective taxes: a note, February 2008.

75 Kruse, Jörn. Internet-Überlast, Netzneutralität und Service-Qualität, Januar 2008.

2007

74 Dewenter, Ralf. Netzneutralität, Dezember 2007

73 Beckmann, Klaus; Gerrits, Carsten. Making sense of corruption: Hobbesian jungle, bribery as an auction, and DUP activities, December 2007.

72 Kruse, Jörn. Crowding-Out bei Überlast im Internet, November 2007.

71 Beckmann, Klaus. Why do petrol prices fluctuate so much?, November 2007.

70 Beckmann, Klaus. Was willst Du armer Teufel geben? - Bemerkungen zum Glück in der Ökonomik, November 2007.

69 Berlemann, Michael; Vogt, Gerit. Kurzfristige Wachstumseffekte von Naturkatastrophen, Eine empirische Analyse der Flutkatastrophe vom August 2002 in Sachsen, November 2007.

68 Schneider, Andrea. Redistributive taxation, inequality, and intergenerational mobility, November 2007.

67 Kruse, Jörn. Exklusive Sportfernsehrechte und Schutzlisten, Oktober 2007.

66 Kruse, Jörn. Das Monopol für demokratische Legitimation und seine Überwindung. Zur konstitutionellen Reform der staatlichen Strukturen, Oktober 2007.

65 Dewenter, Ralf. Crossmediale Fusionen und Meinungsvielfalt: Eine ökonomische Analyse, Oktober 2007.