gender<ed> thoughts

New Perspectives in Gender Research

Working Paper Series
2022, Volume 3

Andrea Klonschinski

Gender Bias in der Wissenschaft

Warum studentische Lehrevaluationen auf den Prüfstand gehören

Mit einem Kommentar von Liudvika Leišytė

GÖTTINGER CENTRUM FÜR GESCHLECHTERFORSCHUNG
GOETTINGEN CENTRE FOR GENDER STUDIES

GÖTTINGER CENTRUM FÜR GESCHLECHTERFORSCHUNG
GOETTINGEN CENTRE FOR GENDER STUDIES

gender<ed>
thoughts

New Perspectives in Gender Research
Working Paper Series

(ISSN 2509-8179)

Official Series of the Göttingen Centre for Gender Studies (GCG)
By 2017 the Göttingen Centre for Gender Studies starts a new working paper series called *Gender(ed) Thoughts Goettingen* as a scholarly platform for discussion and exchange on Gender Studies. The series makes the work of affiliates of the Göttingen Centre visible and allows them to publish preliminary and project-related results.
All contributions to the series will be thoroughly peer-reviewed. Wherever possible, we publish comments to each contribution. The series aims at interdisciplinary exchange among Humanities, Social Sciences as well as Life Sciences and invites researchers to publish their results on Gender Studies. If you would like to comment on existing or future contributions, please get in touch with the editors-in-chief. The series is open to theoretical discussions on established and new approaches in Gender Studies as well as results based on empirical data or case studies. Additionally, the series aims to reflect on Gender as an individual and social perspective in academia and day-to-day life.

2022, Volume 3
Andrea Klonschinski
Gender Bias in der Wissenschaft
Warum studentische Lehrevaluationen auf den Prüfstand gehören

Suggested Citation
Klonschinski, A. (2022): Gender Bias in der Wissenschaft: Warum studentische Lehrevaluationen auf den Prüfstand gehören; Gender(ed) Thoughts, Working Paper Series, Vol. 3, https://dx.doi.org/10.3249/2509-8179-gtg-16 8179-gtg-16 8179-gtg-16 88888179-gtg-21

Göttingen Centre for Gender Studies
Project Office
Georg-August-Universität Göttingen
Centrum für Geschlechterforschung
Platz der Göttinger Sieben 7 • D - 37073 Göttingen
Germany

genderedthoughts@uni-goettingen.de | www.gendered-thoughts.uni-goettingen.de

GÖTTINGER CENTRUM FÜR GESCHLECHTERFORSCHUNG
GOETTINGEN CENTRE FOR GENDER STUDIES

gender<ed> thoughts
New Perspectives in Gender Research
Working Paper Series 2022, Volume 3
DOI: 10.3249/2509-8179-gtg-21

Gender Bias in der Wissenschaft

Warum studentische Lehrevaluationen auf den Prüfstand gehören

Andrea Klonschinski[1]

[1] andrea.klonschinski@gmail.com

Zusammenfassung

Frauen sind unter den Professor:innen nach wie vor unterrepräsentiert. Empirische Befunde deuten darauf hin, dass dies unter anderem auf die systematisch negativeren Bewertungen der Leistung von Frauen in der Wissenschaft zurückzuführen ist. Auch studentische Lehrevaluationen fallen schlechter aus, wenn die Lehrperson eine Frau ist. Während diese Ergebnisse Anlass zum Überdenken der Bewertungspraxis von Wissenschaftler:innen geben sollten, sind studentische Lehrevaluationen nach wie vor die wichtigste und oft einzige an Hochschulen praktizierte Form der Lehrbewertung und ihre Ergebnisse spielen eine wichtige Rolle für die wissenschaftliche Karriere. Der vorliegende Beitrag trägt empirische Befunde zu Gender Bias im Rahmen studentischer Lehrevaluationen zusammen, argumentiert, dass diese Bewertungspraxis ungerecht ist und in ihrer aktuellen Form nicht mehr praktiziert werden sollte und skizziert Vorschläge für alternative, fairere Evaluationsverfahren.

Schlagworte

Studentische Lehrevaluationen, Frauen in der Wissenschaft, Leaky Pipeline, Gender Bias, Stereotype

Abstract

Women are still underrepresented among the higher ranks of academia. Empirical evidence indicates that a central reason for this underrepresentation consists in the fact that women's academic performance is evaluated much more harshly than men's. One example for such evaluation biases are student teaching evaluations, which are systematically worse for female than for male teachers. Considered against the background of the low percentage of female professors, these results should be alarming, but in fact, student teaching evaluations continue to be the major if not the only form of teaching evaluation and their results are of pivotal importance for the academic career. This paper reviews studies on gender biases in student teaching evaluations, argues that they should no longer be used in their current form and sketches fairer alternatives

Keywords

Student Teaching Evaluations, Women in Academia, Leaky Pipeline, Gender Bias, Stereotypes

1. Einleitung

Im Jahr 2020 machen Frauen über alle Fächer hinweg betrachtet die Hälfte der Studierenden, aber nur gut ein Viertel der Professor:innen aus.[1] Und obwohl der Anteil weiblicher Professor:innen in den letzten Jahren ansteigt – 2011 lag er noch bei knapp 19%, 2020 bei 26% – geht es nur langsam voran;[2] die sogenannte *Leaky Pipeline* in der Wissenschaft bleibt undicht. Woran liegt das? Ein Erklärungsansatz rekurriert auf die höheren Ansprüche an und die damit einhergehende strengere und negativere Bewertung der Leistung von Frauen in Führungspositionen im Allgemeinen und in der Wissenschaft im Speziellen (siehe etwa Valian 1998, 2005; Heilman et al. 2004; Heilman/Okimoto 2007; El-Alayli et al. 2018).

Diese Bewertungsverzerrungen dürften nicht zuletzt in dem Spannungsverhältnis zwischen den sozialen Erwartungen an Frauen einerseits und an in der Wissenschaft tätige Personen andererseits begründet sein. Wissenschaftler (hier absichtlich nicht gegendert) werden mit Eigenschaften wie Brillanz, Rationalität, Sachorientierung und Autorität assoziiert (Kaatz et al. 2014: 371; Leslie et al. 2015; Carli et al. 2016; Storage et al. 2016; Krawczyk 2017: 1397; Krawczyk/Smyk 2016: 327). Während diese Charakteristika dem männlichen Genderstereotyp entsprechen, umfasst der Stereotyp für Frauen Attribute wie Fürsorglichkeit, Gemeinschaftsorientierung und Ausdrucksstärke, beinhaltet aber keine professionelle Kompetenz (Valian 1998: 13, 125). Um mit Mary Beard (2017: 53) zu sprechen, *„our mental, cultural template for a powerful person remains resolutely male. If we close our eyes and try to conjure up the image of a president or […] a professor, what most of us see is not a woman.*"[3] Insofern Stereotype nicht nur eine deskriptive, sondern auch eine präskriptive Komponente aufweisen und damit definieren, wie Mitglieder einer bestimmten sozialen Gruppe sein *sollen*, wird die Abweichung von diesen Stereotypen sozial sanktioniert (Valian 1998: 11; Heilman et al. 2004: 416; Dovidio et al. 2010: 7, 8). Das führt in ein Dilemma: Frauen in der Wissenschaft werden entweder bei gleicher Leistung als weniger kompetent beurteilt und weniger respektiert als ihre männlichen Kollegen;[4] oder sie werden nicht gemocht und als unsympathisch, kalt, „bossy" oder unaufrichtig wahrgenommen (Valian 1998: 131; Heilman et al. 2004: 417; Heilman/Okimoto 2007; Bates 2014; Manne 2018: 275).

Mittlerweile deuten zahlreiche empirische Studien darauf hin, dass sich derartige Gender Bias in verschiedenen Bereichen von Academia manifestieren.[5] So konnte etwa gezeigt werden, dass Empfehlungsschreiben für Frauen systematisch negativer ausfallen als die für Männer (Trix/Psenka 2003; Criado-Perez 2019: 102), dass Bewerbungen und wissenschaftliche Texte schlechter bewertet werden, wenn sie von vermeintlich weiblichen Autor:innen stammen (Moss-Racusin et al. 2012; Krawczyk/Smyk 2016) und dass in gemischten Teams erbrachte Leistungen eher den beteiligten Männern als den Frauen zugeschrieben werden (Heilman/Haynes

[1] Für aktuelle Studierendenzahlen siehe https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/11/PD21_538_21.html, Angaben zu den Professor:innen finden sich hier https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/10/PD21_478_213.html, letzter Zugriff 10. Februar 2022.

[2] https://www-genesis.destatis.de/genesis/online?operation=abruftabelleBearbeiten&levelindex=1&levelid=1644478780512&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&code=21341-0001&auswahltext=&werteabruf=starten#abreadcrumb, letzter Zugriff 10. Februar 2022.

[3] Siehe dazu auch Hay (2016).

[4] Eine Analyse der Mimik der Zuhörer:innen konnte etwa negativere Reaktionen gegenüber weiblichen als gegenüber männlichen Führungskräften nachweisen; auch wird letzteren mehr Aufmerksamkeit geschenkt als ersteren (Valian 1998: 130f.).

[5] Die Studien und die folgenden Ausführungen beziehen sich auf Hochschulen in der sogenannten westlichen Welt.

2005; Harson 2016; Wolfers 2016). Als empirisch gut belegt gilt zudem die Tatsache, dass studentische Lehrveranstaltungsevaluationen (LVE) schlechter ausfallen, wenn die Lehrperson eine Frau ist (Sprague/Massoni 2005; Wolbring 2010; Joye/Wilson 2015; MacNell et al. 2015; Miles/House 2015; Wagner et al. 2016; Boring et al. 2016; Boring 2017; Mitchell/Martin 2018; Rivera/Tilcsik 2018; Rosen 2018; Peterson et al. 2019; Mengel et al. 2019; Murray et al. 2020; Boring/Philippe 2021; Kreitzer/Sweet-Cushman 2021). Nichtsdestotrotz sind studentische Lehrevaluationen mittels standardisierter Fragebögen auch an deutschen Hochschulen weit verbreitet und stellen häufig die einzige praktizierte Form der Lehrevaluation dar. Insofern ihre Ergebnisse für die Vergabe von Lehrpreisen, zur internen Kontrolle der Lehrleistung sowie in Bewerbungs- und Berufungsverfahren eine wichtige Rolle spielen (DGS 2020), haben sie eine erhebliche Bedeutung für die wissenschaftliche Karriere. Diese Praxis, so die These des vorliegenden Beitrags, führt zu einer ungerechtfertigten Benachteiligung von Frauen (und anderen marginalisierten Gruppen) in Academia. Da zudem strittig ist, inwiefern studentische LVE überhaupt Auskunft über die Qualität der Lehre liefern (z. B. Staufenbiel et al. 2015; Boring et al. 2016; Uttl et al. 2017; Hornstein 2017), sollten sie in ihrer derzeitigen Form nicht mehr praktiziert werden.

Um diese Forderung zu untermauern, werden im Folgenden zentrale empirische Ergebnisse zu genderspezifischen Verzerrungen in studentischen LVE vorgestellt (Abschnitt 2). Dabei wird sich zeigen, dass die Lehre von Frauen nicht nur systematisch negativer bewertet wird als die ihrer männlichen Kollegen, sondern anhand anderer, genderspezifischer Bewertungsmaßstäbe erfolgt, welche wesentlich schwieriger zu erfüllen sind. Auf Basis dieser Darstellung wird herausgearbeitet, inwiefern studentische LVE Ungerechtigkeiten verstärken und skizziert, wie eine fairere Evaluation der Hochschullehre aussehen könnte (Abschnitt 3). Das Fazit fasst die Überlegungen zusammen (Abschnitt 4).

Die hier zusammengetragenen Ergebnisse und die Argumentation sind nicht neu.[6] Die Relevanz des Beitrags ergibt sich vielmehr daraus, dass Gender Bias in Lehrevaluationen in der Deutschen hochschulpolitischen Debatte und der Praxis bislang zu wenig Beachtung finden.[7] Der Beitrag zielt vor diesem Hintergrund darauf ab, die dargestellten Studienergebnisse disziplinenübergreifend einem breiteren Leser:innenkreis zugänglich zu machen und eine Debatte über den Umgang mit studentischen LVE anzustoßen. Es handelt sich daher nicht um einen umfassenden systematischen Literaturreview. Die Auswahl der vorgestellten Beiträge orientiert sich an ihrer wahrgenommenen Relevanz für die Debatte und an ihrer Aktualität (nach 2000), da die Validität der Ergebnisse älterer Untersuchungen mit dem Verweis auf veränderte Einstellungen zu Gleichberechtigung, der Zunahme des Frauenanteils in der Wissenschaft u. ä. in Frage gestellt werden kann. Zudem fokussiert sich die Darstellung auf *gender*bedingte Bewertungsverzerrungen und die Benachteiligung von Frauen. Damit soll nicht in Abrede gestellt werden, dass nicht-binäre Personen sowie Mitglieder anderer (in der Wissenschaft) marginalisierter Gruppen wie Erstakademiker:innen oder Menschen mit Migrationshintergrund durch studentische LVE benachteiligt werden und sich verschiedene Dimensionen der Benachteiligung wechselseitig verstärken können (Chisadza et al. 2019; Fan et al. 2019; Chávez/Mitchell 2020). Die Auseinandersetzung mit Benachteiligungen aufgrund der Kategorie Gender soll auch dazu dienen, die Sensibilität für andere Dimensionen der Benachteiligungen zu schärfen zumal, wie im Folgenden deutlich werden wird, die der verzerrten Bewertungen zugrundeliegenden sozial-

[6] Für aktuelle Reviews zu Gender Bias in studentischen LVE siehe etwa Heffernan (2021) und Kreitzer/Sweet-Cushman (2021).

[7] Eine Ausnahme ist die Stellungnahme der Deutschen Gesellschaft für Soziologie zum Umgang mit studentischen Lehrveranstaltungsevaluationen (DGS 2020).

psychologischen Mechanismen in Bezug auf verschiedene marginalisierte Gruppen einander ähneln.[8]

2. Gender Bias in Studentischen Lehrevaluationen

2. 1 Studentische Lehrevaluationen und Biasvariablen

Die moderne Form studentischer LVE bildete sich in der Mitte des 20. Jahrhunderts heraus und ist seither Gegenstand von Kritik (Stark/Freishtat 2014). Im Zentrum stehen dabei insbesondere folgende Fragen: Was genau *sollen* die Evaluationen messen (den Lernerfolg der Studierenden? Ihre Zufriedenheit? Die Qualität der Lehre?) und wie ist diese Zielgröße zu definieren bzw. zu operationalisieren (lässt sich zum Beispiel Lernerfolg an Noten ablesen?). Was messen studentische LVE *tatsächlich* und wie hängen diese Daten mit der definierten Zielgröße zusammen? Hinsichtlich des Zusammenhangs von studentischen Lehrevaluationen und Lernerfolg der Studierenden kommen Bob Utt et al. (2017: 40) in einer Meta-Studie zu einem ernüchternden Ergebnis: *„The best evidence […] indicates that the SET [student teaching evaluation] / learning correlation is zero"*; mit anderen Worten: *„students do not learn more from professors who receive higher SET ratings"*. Die im aktuellen Umgang mit studentischen LVE präsumierte Unterstellung, dass diese Auskunft darüber geben, wie „gut" eine Person als Lehrende ist, ist damit mehr als fraglich (DGS 2020).

Tatsächlich werden studentische LVE von einer Vielzahl von Faktoren beeinflusst, die die Lehrperson gar nicht oder nur in begrenztem Ausmaß beeinflussen kann sowie von Einflussgrößen, die sie zwar kontrollieren kann, die aber in keinem Zusammenhang mit der Qualität der Lehre stehen. In die erste Kategorie fallen neben den hier interessierenden Charakteristika der Lehrperson wie Gender und Alter etwa die Größe des Kurses, die Tatsache, ob es sich um einen Wahl- oder Pflichtkurs handelt oder das ex ante Interesse der Studierenden am Kursinhalt (Miles/House 2015; Staufenbier et al. 2015; Murray et al. 2020; Kreitzer/Sweet-Cushman 2021). In die zweite Kategorie fällt zum Beispiel eine besondere Milde bei der Notengebung (Staufenbiel et al. 2015: 45). Staufenbiel et al. (ebd.) weisen in diesem Kontext auf die Schwierigkeit hin, ohne empirische Daten zu bestimmen, welche Faktoren irrelevant für die Beurteilung der Qualität der Lehre sind. So sei etwa denkbar, dass Studierende in großen Gruppen weniger lernen. Analog ist nicht *prinzipiell* auszuschließen, dass männliche, weiße, heterosexuelle, akzentfrei sprechende, attraktive, mitteljunge Professoren (Murray 2020; Heffernan 2021: 5) tatsächlich die besten Lehrer sind. Die Studienlage deutet allerdings stark darauf hin, dass dies nicht der Fall ist und Gender vielmehr tatsächlich eine urteilsverzerrende Biasvariable darstellt. Dies soll im Folgenden anhand der detaillierteren Darstellung zentraler Studien illustriert werden. Dabei werden zunächst Untersuchungen dargestellt, die eine *schlechtere* Bewertung weiblicher Lehrender aufdecken, um in einem zweiten Schritt auf Studienergebnisse einzugehen, die nahelegen, dass Frauen dabei anhand *anderer* Kriterien bewertet werden als ihre männlichen Kollegen.

2. 2 Negativere *Bewertung von Frauen*

Da sich Lehrende in Persönlichkeit und Lehrstil unterscheiden, ist es schwierig, den spezifischen Einfluss einzelner Faktoren in einer komplexen Lehrsituation zu eruieren. Deshalb ist das Experiment von Lillian MacNell et al. (2015) hinsichtlich möglicher Einflüsse der Kategorie Gender besonders aufschlussreich, obwohl die Anzahl der Versuchspersonen mit 43 sehr gering ist. Die Autor:innen der Studie betrachteten einen

[8] Ich danke zwei anonymen Gutachter:innen für den Hinweis auf die nötige Klärung dieser Punkte sowie für wertvolle Anmerkungen zum Manuskript.

Einführungskurs in Anthropologie/Soziologie an der Universität von North Carolina. Dieser Kurs fand online statt und die Teilnehmer:innen hatten lediglich digitalen Kontakt zu den Dozent:innen. Geleitet wurde der Kurs von einem Professor, der Aufgaben und Kursmaterial verteilte; die hauptsächlichen Interaktionen zwischen Student:innen und Lehrpersonen fanden indes in sechs Diskussionsgruppen statt, deren Leiter:innen eine Note vergaben. Zwei dieser Gruppen wurden vom leitenden Professor übernommen, jeweils zwei weitere von einer weiblichen und einer männlichen Lehrkraft, nennen wir sie Martin und Martina, die jeweils einen Kurs unter ihrer eigenen Identität unterrichteten und einen unter der des bzw. der Kollegin. Martin unterrichtete also einen Kurs als Martin und einen als „Martina". Dabei wurde Sorge getragen, dass Lehre und Verhalten der beiden Dozent:innen einander möglichst glichen (MacNell et al. 2015: 296). Die Studierenden wurden gebeten, die Kurse anschließend anhand verschiedener Kriterien zu bewerten.

Im Ergebnis gab es keine relevanten Unterschiede zwischen den Bewertungen der beiden tatsächlich von Martin und Martina geleiteten Kurse. Ein Vergleich der beiden nominell von „Martin" bzw. „Martina" geleiteten Kursevaluationen zeigte jedoch, dass die als männlich identifizierte Lehrperson hinsichtlich aller Kriterien (Professionalität, Schnelligkeit, respektvolles Verhalten, Enthusiasmus, Lob, Fürsorglichkeit, Konsistenz, Fairness, Feedback, Hilfsbereitschaft, Ansprechbarkeit besser bewertet worden ist als die weibliche, zum Teil signifikant (MacNell et al. 2015: 298). Besonders interessant ist die unterschiedliche Bewertung der Schnelligkeit, da die Lehrpersonen sich miteinander abgestimmt und Feedback tatsächlich gleich schnell gegeben haben. Wenn „Martin" und „Martina" Noten nach zwei Tagen veröffentlicht haben, lieferte das für „Martin" nichtsdestotrotz eine Bewertung von durchschnittlich 4,35 von 5 für Schnelligkeit, während „Martina" nur einen Wert von 3,55 erzielte (ebd.: 300).

Die Ergebnisse von MacNell et al. wurden in einer Untersuchung von Kristina Mitchell und Jonathan Martin bestätigt (Mitchell/Martin 2018). Beide hatten im Frühjahr 2015 mehrere Online-Einführungskurse in Politischer Theorie an einer US-Amerikanischen Universität gehalten. Die Kurse glichen sich vollständig in Aufbau, Inhalt und Prüfungsleistung und unterschieden sich lediglich hinsichtlich des bzw. der Kursleiter:in sowie den Personen, die die Prüfungsleistungen benoteten (ebd.: 650). Die studentische LVE fragte nach der Bewertung von fünf Aspekten: Kursleiter:in (z. B. Fairness und Anregung), Kursleiter:in/Kurs (z. B. die Fähigkeit, Inhalte zu vermitteln), Seminar (z. B. Erwartungen und Arbeitspensum), Technik (Bereitstellung von Informationen u. ä.) und Verwaltung (Anmeldung u. ä.). Auf den letzten Aspekt hatten die Kursleiter:innen keinen Einfluss. Auch hier stellte sich heraus, dass die Kurse von Kristina Mitchell hinsichtlich aller Aspekte – außer Verwaltung – schlechter bewertet wurden als die von Jonathan Martin: *„The data are clear: a man received higher evaluations in identical courses, even for questions unrelated to the individual instructor's ability, demeanor, or attitude"* (ebd.: 651).

Auf negativere Bewertungen von weiblichen Lehrpersonen weisen auch die in Europa durchgeführten Studien von Anne Boring (2017) und Friederike Mengel et al. (2019) hin. Mengel et al. (2019) testen den Einfluss des Geschlechts der Lehrkräfte auf die LVE im Rahmen eines natürlichen Experiments an der *School of Business and Economics* der Maastricht Universität. Ein solches Experiment ist möglich, weil die Studierenden zufällig einem Kurs mit männlicher oder weiblicher Lehrperson zugeordnet werden und am Ende des Semesters dieselbe Prüfung ablegen. Es finden somit keine Verzerrungen durch Selbstselektion der Evaluierenden statt und der Lehr- bzw. Lernerfolg kann am Ende des Semesters anhand des Prüfungsergebnisses über verschiedene Kurse und somit Dozent:innen hinweg verglichen werden (so auch Boring 2017: 28f.). Es zeigt sich, dass die Studierenden weibliche Lehrpersonen signifikant schlechter bewerten (Mengel et al. 2019: 552f.). Männliche Studierende bewerten zudem auch Faktoren wie das Funktionieren der Gruppe sowie das Lehrmaterial schlechter, wenn die Lehrperson weiblich ist – und das, obwohl das Lehrmaterial in

verschiedenen Kursen identisch war (ebd.: 554f.). Dabei hat das Geschlecht der Lehrperson keinen Einfluss auf aktuelle und zukünftige Noten oder den berichteten Arbeitsaufwand der Studierenden, woraus die Autor:innen der Studie schließen, dass sich die Bewertungen nicht auf unterschiedliche objektive Leistungen der Lehrperson zurückführen lassen (ebd.: 557).[9] Eine Analyse nach Statusgruppe der Lehrenden zeigt, dass Doktorandinnen sowohl bei Studenten als auch bei Studentinnen schlechter abschneiden als ihre männlichen Kollegen, wohingegen sich der Unterschied auf Ebene der Professor:innen einebnet (ebd.: 557f.). Dies ist im Hinblick auf das Problem der *Leaky Pipeline* besonders problematisch.

Boring (2017) testet den Einfluss des Geschlechts der Lehrkräfte auf die LVE im Rahmen eines natürlichen Experiments an einer französischen Universität.[10] Im Rahmen der Lehrevaluation von sechs Pflichtkursen[11] sollen die Studierenden die folgenden Dimensionen bewerten:

„Course content: the professor's preparation and organization of classes, and the quality of instructional materials. Assignments: the clarity of the assessment criteria, and usefulness of feedback. Delivery style: ability to lead the class, ability to encourage group work, and the professor's availability and quality of personal contact. Professor's knowledge: the ability to relate to current issues, and the professor's contribution to the student's intellectual development" (Boring 2017: 29).

Schließlich werden sie nach ihrer allgemeinen Zufriedenheit gefragt. Im Ergebnis zeigte sich kein Unterschied in den Noten der von weiblichen oder männlichen Lehrpersonen unterrichteten Studierenden. Während indes männliche Dozenten signifikant höhere allgemeine Zufriedenheitswerte von männlichen und weiblichen Studierenden bekamen (ebd.: 31), ist der Zusammenhang von Gender und Bewertung hinsichtlich der einzelnen Dimensionen komplexer. Die Unterschiede waren am größten, wenn die Dimension eine große Nähe zu einem Genderstereotyp aufweist, wie es etwa bei „ability to lead the class" oder den unter „Professor's knowledge" versammelten Faktoren und dem männlichen Genderstereotyp der Fall ist (ebd.: 33f.). Insofern die beobachteten Gender Bias aber je nach bewerteter Dimension unterschiedlich ausfallen, gebe es keine Möglichkeit, die Evaluation *ex post* entsprechend zu korrigieren, indem der Bias wieder „herausgerechnet" würde (ebd.: 36).

2. 3 Andere *Bewertung von Frauen*

Die bisher dargestellten Studien deuten darauf hin, dass es sich bei Gender um eine Biasvariable handelt, die studentische LVE negativ beeinflusst, aber nicht mit der Qualität der Lehre zusammenhängt. Mit den in den jeweiligen Stereotypen gebündelten konfligierenden Erwartungen an „Frauen" und „Wissenschaftler" bzw. Personen in Führungspositionen wurde eingangs bereits ein Erklärungsansatz für diese negativeren Bewertungen skizziert. Auf die zentrale Rolle von Stereotypen hebt auch Boring (2017: 35) in der Zusammenfassung ihrer Ergebnisse ab:

„Overall, these results suggest that gender stereotypes may be driving students' evaluations of professors. Students sometimes reward (or at least do not penalize) women on stereotypically female criteria, while systematically rewarding men on stereotypically male criteria."

Tatsächlich deutet einiges darauf hin, dass Studierende genderspezifische Erwartungen an das Lehrpersonal hegen. Ausgehend von der Überlegung, dass sich Stereotype und damit genderspezifische Bewertungsmaßstäbe in der Art und

[9] In einer Studie von Shauna W. Joye und Janie H. Wilson (2015) zeigte sich, dass Versuchspersonen verbal vermittelte Informationen anscheinend besser memorieren, wenn die Dozentin eine Frau über 40 ist, obwohl diese „älteren" Frauen als weniger attraktiv als jüngere Frauen und Frauen insgesamt als weniger effektiv als ihre männlichen Kollegen bewertet wurden.

[10] Wie bei Mengel et al. (2019) können sich die Studierenden die Kurse nicht nach Dozent:in aussuchen, müssen eine Lehrevaluation durchführen und absolvieren dieselbe Prüfung.

[11] Es handelt sich um Einführung in die Mikroökonomik, Politische Institutionen und Geschichte während des Herbstsemesters sowie Einführung in die Makroökonomik, Politikwissenschaft und Soziologie im Frühjahrssemester (Boring 2017: 28).

Weise, wie über Lehrende gesprochen wird, manifestieren, baten Joey Sprague und Kelley Massoni (2005) 288 Studierende, im Rahmen einer schriftlichen Befragung die beste sowie die schlechteste Lehrperson zu beschreiben, die sie jemals hatten. Die dabei verwendeten Adjektive wurden von Sprague und Massoni zu mehreren Synonym-Clustern (z. B. *brillant, smart, knowledgeable*) verdichtet und diese wiederum in mehrere Dimensionen (hier etwa: intelligent) gruppiert. Dabei zeigten sich genderspezifische Unterschiede. Während etwa sowohl Lehrer als auch Lehrerinnen mit Eigenschaften versehen werden, die unter die Dimension „*nurturing*" fallen, sind diese Charakterisierungen bei der Beschreibung von Frauen häufiger und vielfältiger. Bestimmte Cluster-Begriffe, wie *giving* und *compassionate* wurden ausschließlich zur Beschreibung weiblicher Lehrkräfte verwendet (ebd.: 786). Schlechte Lehrkräfte wurden unabhängig vom Geschlecht als langweilig bezeichnet; schlechte Lehrer zudem als *rude* und *arrogant*, Lehrerinnen als *mean* und *unfair*. Wiederum finden sich einige stark gegenderte Beschreibungen, etwa innerhalb der Dimensionen *rude* und *mean* (ebd.: 788). In beiden Fällen ist die Varianz der verwendeten Beschreibungen von Frauen wesentlich größer als die für Männer. Es zeigt sich zudem, dass die Begriffe zur Beschreibung von Frauen respektloser sind als die für Männer (z. B. *insolent, phony, fake, bitchy, witch, feminazi*); „the most hostile words are saved for women teachers" (ebd.: 791).[12] Tatsächlich scheinen weibliche Lehrkräfte häufiger mit respektlosem Verhalten der Studierenden konfrontiert zu sein als Männer (Saul 2013; Hay 2016; Eidinger 2017; Manne 2018). Sie werden öfter um Verlängerungen, Notenverbesserungen oder andere Nachsichtigkeiten gebeten als ihre männlichen Kollegen und die Reaktionen der Studierenden fallen negativer aus, wenn ihre Bitten erfolglos bleiben (El-Alayli et al. 2018). Diese Befunde reihen sich ein in die eingangs genannte Beobachtung, dass Frauen größere Probleme haben, als Autoritäten anerkannt zu werden.

In der abschließenden Interpretation ihrer semantischen Analyse stellen Sprague und Massoni (ebd.: 791) hinsichtlich gegenderter Bewertungsstandards fest:

> *„Men teachers are more likely to be held up to an entertainer standard: are they funny, or is their performance a failure because that [sic] are arrogant and bore their audiences? Women teachers are held to a nurturer standard: are they caring and nurturing, or are their relationships with the students a failure because they are mean, unwilling to negotiate (rigid, unfair), or hard to relate to (cold psychotic)?"*

Es ist zu betonen, dass der *nurturer* Standard wesentlich schwerer zu erfüllen ist als der des Entertainers. Während die unterhaltsame Vorlesung, zugespitzt formuliert, jedes Jahr wiederholt werden kann, erfordert die ständige Ansprechbarkeit und die, auch emotionale, Betreuung von Studierenden viel Zeit und Energie (ebd.). Dass weibliche Lehrkräfte in der Tat mehr Zeit für Lehre und Betreuung von Studierenden sowie für sonstige *Care*- und Service-Tätigkeiten innerhalb Academia aufwenden, kann als gut belegt gelten (Linka et al. 2008; Misra et al. 2011; Saul 2013; Hay 2016; Boring 2017; El-Alayli et al. 2018: 148; Criado-Perez 2019: 97f.; Felkey/Batz-Barbarich 2021).[13] Diese Tätigkeiten bleiben allerdings meist unsichtbar, sind wenig

[12] So auch Kreitzer/Sweet-Cushman (2021). Derartige Unterschiede lassen sich mithilfe eines Tools von Benjamin Schmidt auch graphisch illustrieren. Schmidt hat über 14 Millionen Studierendenevaluationen von ratemyprofessor.com analysiert und die Häufigkeit der Verwendung bestimmter Begriffe je nach akademischer Disziplin und nach Geschlecht aufgeschlüsselt. Das interaktive Tool erlaubt die Eingabe von Begriffen und liefert die entsprechenden in einem Graphen aufbereiteten Daten (https://bit.ly/3BqQvRR, letzter Zugriff 16. Februar 2021. Manne (2018: 275) weist darauf hin, dass der Begriff „fake" insgesamt, wenn auch nicht in jeder Disziplin, wesentlich häufiger zur Beschreibung von Frauen als von Männern verwendet wird und vermutet „that people are more inclined to see women in positions of authority as posers and imposters compared with their male counterparts".

[13] Joya Misra et al. (2011) berichten: „A variety of studies show that men focus more on research than do women. While men are not necessarily more productive than women, they are more protective of their research time. Tenured women, on the other hand, devote much more time to teaching, mentoring, and service, and particularly to activities that may be seen as building bridges around the university."

prestigeträchtig und damit der akademischen Karriere nicht unbedingt förderlich.[14]

3. Schlussfolgerungen und Maßnahmen

Die hier referierten Studien deuten auf die Existenz von Gender Bias in studentischen LVE hin. Nicht nur wird die Lehre von Frauen hinsichtlich derselben Dimensionen (z. B. Kompetenz, Schnelligkeit des Feedbacks und Lehrmaterial) schlechter bewertet, sie wird auch anhand anderer Maßstäbe evaluiert (*Care* vs. *Entertainment*). Was folgt daraus? Zunächst ließe sich einwenden, dass die messbaren Effekte von Gender Bias zu geringfügig sind (Staufenbiel et al. 2015; Rosen 2018), um zu einer spürbaren Benachteiligung von Frauen zu führen und den hier dargestellten Ergebnissen daher keine praktische Relevanz zukommt. Diesem Einwand ist entgegenzuhalten, dass sich die systematisch[15] schlechteren Ergebnisse studentischer LVE einreihen in systematisch negativere Bewertungen von Frauen in der Wissenschaft generell. Es steht zu vermuten, dass geringfügige Effekte sich über verschiedene Kontexte und die Zeit hinweg kumuliert in dem hoch kompetitiven Wissenschaftssystem in substanziellen Nachteilen für Frauen niederschlagen (Valian 1998, 2005; Brennan 2013; Chávez/Mitchell 2019: 271). Es gilt zudem zu betonen, dass zu der hier betrachteten Benachteiligung anhand der Kategorie Gender Benachteiligungen entlang anderer Gruppenmerkmale, wie Ethnizität, Muttersprache oder wahrgenommene sexuelle Orientierung hinzutreten und sich wechselseitig verstärken dürften (Heffernan 2021: 8). Auch dürften die negativen Effekte für Frauen und andere marginalisierte Gruppen besonders prononciert in denjenigen (insbesondere quantitativen) Disziplinen ausfallen, die stark mit stereotyp „männlichen" Eigenschaften assoziiert werden und zahlenmäßig nach wie vor von Männern dominiert werden, wie etwa die MINT-Fächer (Mathematik, Informatik, Naturwissenschaft und Technik) oder die Philosophie (Leslie et al. 2015; Klonschinski 2018, 2020; Felkey/Batz-Barbarich 2021).

Hinsichtlich der praktischen Konsequenzen dieser unfairen Benachteiligung scheint zunächst die Forderung naheliegend, studentische LVE nicht mehr zur Grundlage der externen Bewertung von Lehrkräften zu machen, um Frauen etwa bei der Besetzung von Stellen nicht zu benachteiligen. Obwohl ein Schritt in die richtige Richtung, wäre diese Konsequenz nicht weitreichend genug. So ist angesichts der Erkenntnis, dass studentische Lehrevaluationen genderbedingten Verzerrungen unterliegen, unklar, warum sie in ihrer derzeitigen Form intern als Monitoring-Instrument der Lehre oder auch „nur" als Feedback für die Lehrperson überhaupt herangezogen werden sollen. Auch diese Verwendungen können sich negativ auf Frauen auswirken. So könnten sie in Antizipation der höheren studentischen Erwartungen sowie mittelmäßiger Evaluationen mehr Zeit in die Lehre investieren, was (noch) weniger Zeit für Forschung impliziert. Auch steht zu erwarten, dass schlechte Evaluationen und insbesondere respektlose Bemerkungen gerade für junge Dozentinnen mit einem hohen Maß an Frustration verbunden sind, sie ihnen die Freude am Lehren und somit letztlich die Lust auf die universitäre Karriere nehmen.[16] Da zudem der Zusammenhang zwischen studentischen LVE und der Qualität der Lehre fraglich ist und studentische LVE falsche Anreize bei der Gestaltung von Lehre und Prüfungen setzen (DGS 2020), sollten sie in ihrer traditionellen standardisierten Form nicht mehr praktiziert und verwendet werden (Stark/Freishtat 2014; Mitchell/Martin 2019;

[14] Jennifer Saul (2013: 45) schreibt dazu: „Because women are more associated than men with interpersonal and helping skills, they're likely to be assigned more of the time-intensive student support and administrative/service tasks that tend to be poorly rewarded in terms of promotion."

[15] So stellt Rosen (2018: 42) fest: „it still appears that women are at a particular disadvantage when it comes to student evaluations, as there are no disciplines where women have statistically higher *overall quality* scores".

[16] Auch wenn es scheint, dass nur eine kleine Minderheit der Studierendenkommentare beleidigend sind (Tucker 2014), ist anzunehmen, dass diese einen besonders bleibenden Einfluss auf die Lehrperson ausüben (Himelein 2017).

Chavez/Mitchell 2019; DGS 2020; Heffernan 2021; Kreitzer/Sweet-Cushman 2021).

Das heißt nicht, dass auf die Evaluation von Lehre verzichtet werden oder dass keinerlei Rückmeldung von Studierenden eingeholt werden sollte. Alternative Methoden zur Evaluation der Lehre sollten jedoch holistischer erfolgen, als es bisher an den meisten Universitäten der Fall ist (Stark/Freishtat 2014; Falkoff 2018). Eine holistische(re) Evaluation der Lehre begreift studentisches Feedback zu Lehrveranstaltungen als *einen* Bestandteil einer umfassenderen Bewertungspraxis. Studentische Rückmeldungen sind dabei nicht als letztgültige Urteile über die Qualität oder Effektivität der Lehre zu verstehen, sondern als Informationen über ihr Erleben des Kurses (ebd.; Hornstein 2017; Fan et al. 2019; Kreitzer/Sweet-Cushman 2021). Dabei sind sowohl subjektive Einschätzungen zum eigenen Lernerfolg und Interesse als auch Antworten auf deskriptive Fragen, wie etwa, wie lange sie im Schnitt auf Antworten auf ihre E-Mails gewartet oder wie viel Zeit sie auf die Vorbereitung des Kurses verwendet haben. Flankiert werden sollten derartige studentische LVE von weiteren Evaluationsmethoden. Kolleg:innen, Vorgesetzte oder externe Gutachter:innen könnten sich etwa regelmäßig Seminarpläne und sonstige Kursmaterialien ansehen, in Lehrveranstaltungen hospitieren, nach der Selbsteinschätzung der Lehrperson fragen und diese Daten mit studentischen Aussagen abgleichen.[17] Flankiert werden sollte dieser Evaluierungsprozess von Fortbildungsmöglichkeiten und Unterstützungsangeboten, wie etwa Peer-Mentoring-Programmen (DSG 2020: 3).

Insofern all diese Verfahren ebenfalls anfällig für Gender Bias sind, gilt es, bei der Gestaltung von Evaluationsmaßnahmen auf die Reduzierung des Einflusses von Genderstereotypen hinzuwirken. So zeigen etwa Studien von David Peterson et al. (2019) sowie von Anne Boring und Arnaud Philippe (2021), dass die Information über das Vorliegen von Gender Bias in studentischen LVE die Bewertung weiblicher Lehrkräfte verbessert. Ein bloßer Appell, im Rahmen der LVE nicht zu diskriminieren, blieb indes wirkungslos. Angesichts der bislang überschaubaren Forschungslage zur Reduktion von Bias in studentischen LVE (Kreitzer/Sweet-Cushman 2021) sowie in Anbetracht der Tatsache, dass es bei derartigen Maßnahmen auch zu unerwünschten Rebound-Effekten kommen kann (Dobbin/Kalev 2018) ist hier weitere Forschung vonnöten. Dasselbe gilt für die Frage, wie sich das Evaluationsdesign auf die Aktivierung von Stereotypen auswirkt. Um zwei Beispiele zu nennen: Einer Studie von Lauren A. Rivera und András Tilcsik (2018) zufolge wirkt sich etwa die Wahl der Skala auf das Ausmaß genderspezifischer Bewertungsverzerrungen aus. So sei es für Frauen wesentlich wahrscheinlicher, auf einer Skala von 0 bis 6 mit einer 6 bewertet zu werden, als auf einer Skala von 0 bis 10 mit einer 10. Iris Bohnet et al. (2016) wiederum zeigen, dass die simultane Bewertung von mehreren Bewerber:innen unterschiedlichen Geschlechts den Einfluss von Gender Bias nivelliert und die Leistung der Bewerber:innen in den Vordergrund treten lässt. Die weitere Erforschung derartiger Effekte der Architektur von Lehrevaluationen scheint ein wichtiger Schritt hin zu *Gender Equality by Design* (Bohnet 2018) in Academia.

4. Fazit

Empirische Studien deuten darauf hin, dass die Leistung von Frauen in der Wissenschaft Gegenstand genderspezifischer Bewertungsverzerrungen ist, sodass diese Tatsache ein Grund für die nach wie vor undichte *Leaky Pipeline* in Academia sein dürfte. Der vorliegende Beitrag hat empirische Ergebnisse zu Gender Bias in studentischen LVE zusammengetragen, um dieses Thema auf die Agenda der deutschsprachigen hochschulpolitischen Debatte zu setzen. Den Studien zufolge werden Frauen im Rahmen studentischer LVE nicht nur insgesamt schlechter bewertet als ihre

[17] Ein Beispiel für ein derartiges Vorgehen liefern Stark und Freishtat (2014: 5f.). Peer-Review-Verfahren schlagen auch Wagner et al. (2016) Kreitzer/Sweet-Cushman (2021) vor, Peer-Mentoring fordert DSG (2020).

männlichen Kollegen; sie werden auch anhand anderer, schwerer zu erfüllender und spezifisch gegenderter Maßstäbe beurteilt. Da systematisch negativere Evaluationsergebnisse nicht nur zu Nachteilen bei der Stellenvergabe führen, sondern sich auch negativ auf die Motivation auswirken und die Arbeitslast von Frauen im Bereich *Care*- und *Service*-Arbeit in Academia weiter verstärken dürften, sollten sie in ihrer derzeitigen Form nicht mehr durchgeführt werden. Die herausgehobene Rolle, die studentischen LVE derzeit an den Hochschulen zukommt, steht den Bekenntnissen der Institutionen zu Diversität und Nicht-Diskriminierung diametral gegenüber.[18] Als Alternativen wurden holistischere Verfahren mit *Peer-Review*- und *Peer-Mentoring*-Elementen ausgemacht. Zudem wurde auf Forschungsbedarf hinsichtlich der Effekte der Evaluationsarchitektur auf Gender Bias hingewiesen. Weitere Forschung ist ebenfalls im Hinblick auf andere Dimensionen der Benachteiligung, wie etwa Migrationshintergrund oder Behinderung, vonnöten.

Literatur

Bates, Laura. 2014.: „When will we stop calling successful women 'abrasive'?" In: *The Guardian*, 3. Oktober. Letzter Zugriff am 15. Februar 2022. https://www.theguardian.com/lifeandstyle/womens-blog/2014/oct/03/when-will-we-stop-calling-successful-women-abrasive

Beard, Mary. 2017. *Women and Power. A Manifesto.* London: Profile Books.

Bohnet, Iris. 2016. *What works. Gender Equality by Design.* Cambridge und London: Belknap Press.

Bohnet, Iris; Van Geen, Alexandra und Max Bazerman. 2015. „When performance trumps gender bias: joint vs separate evaluation." In: *Management Science* 62, 5: 1225-1234.

Boring, Anne. 2017. „Gender biases in student evaluations of teaching". In: *Journal of Public Economics* 145: 27-41.

Boring, Anne; Ottoboni, Kellie und Philip B. Stark. 2016. „Student evaluations of teaching (mostly) do not measure teaching effectiveness". In: *Science Open Research.* DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Boring, Anne und Arnaud Philippe. 2019. „Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching." In: *Journal of Public Economics* 193, 104323.

Brennan, Samantha. 2013. „Rethinking the relevance of micro-inequities: the case of women in philosophy." In: Hutchison, Katrine und Fiona Jenkins (Hg.): *Women in Philosophy. What Needs to Change?* 180-196. Oxford u. a.: Oxford University Press.

Carli, Linda L.; Alawa, Laila; Lee, YoonAh; Zhao, Bei und Elaine Kim. 2016. „Stereotypes about gender and science: women ≠ scientists". In: *Psychology of Women Quarterly* 40, 2: 244-260.

Chávez, Kerry und Kristina M.W. Mitchell. 2020. „Exploring bias in student evaluations: gender, race, and ethnicity". In: *PS: Political Science & Politics* 53, 2: 270–274.

Chisadza, Carolyn, Nicholls, Nicky und Eleni Yitbarek. 2019. „Race and gender biases in student evaluations of teaching". In: *Economic Letters* 179: 66-71.

Criado-Perez, Carolina 2019: *Invisible Women: Exposing Data Bias in a World designed for Men.* London: Chatto & Windus.

[18] Troy Heffernan (2021: 9) formuliert es deutlich: „no university […] can declare to be a gender equal employer or have an interest in growing a safe, inclusive and diverse workforce if they continue using SETs to evaluate course and teacher quality".

DGS. 2020. „Stellungnahme der Deutschen Gesellschaft für Soziologie zum Umgang mit Studentischen LVE". Letzter Zugriff 24. August 2022. https://soziologie.de/fileadmin/user_upload/stellungnahmen/DGS-Stellungnahme_Lehrveranstaltungsevaluation_31.08.2020.pdf.

Dobbin, Frank und Alexandra Kalev. 2018. „Why doesn't diversity training work? The challenge for industry and academia". In: *Anthropol. Now* 10, 2: 48-55.

Dovidio, John F.; Hewstone, Miles; Glick, Peter, und Victoria M. Esses. 2010. „Prejudice, stereotyping and discrimination: theoretical and empirical overview". In: Dies. (Hg.): *The SAGE Handbook of Prejudice, Stereotyping and Discrimination:* 3-28. Los Angeles: SAGE.

Eidinger, Andrea. 2017. „She's hot: female sessional instructors, gender bias, and student evaluations". *Active History*. 30. März. Letzter Zugriff 16. Februar 2022. http://activehistory.ca/2017/03/shes-hot-female-sessional-instructors-gender-bias-and-student-evaluations/.

El-Alayli, Amani; Hansen-Brown, Ashley A. und Michelle Ceynar. 2018. „Dancing backwards in high heels: Female professors experience more work demands and special favor requests, particularly from academically entitled students". In: *Sex Roles* 79, 3-4: 136-150.

Falkoff, Michelle. 2018. „Why we must stop relying on student ratings of teaching". *The Chronicle of Higher Education*. Letzter Zugriff am 16. Februar 2022.https://www.chronicle.com/article/Why-We-Must-Stop-Relying-on/243213

Fan, Yanan; Shepherd, Laura J.; Slavich, Eve; Waters, Donna; Stone, M.; Abel, R. und Emma L. Johnson. 2019. „Gender and cultural bias in student evaluations: Why representation matters". *PLoS ONE* 14, 2: e0209749. https://doi.org/10.1371/journal.pone.0209749.

Felkey, Amanda J. und Cassondra Batz-Barbarich. 2021. „Can women teach math (and be promoted)? A meta-analysis of gender differences across student evaluations of teaching." In: *AEA Papers and Proceedings* 111: 184-189.

Hay, Carol. 2016: „Girlfriend, mother, professor?" *New York Times* vom 25. Januar. Letzter Zugriff 16. Februar 2022. https://www.uml.edu/news/news-articles/2016/nyt-girlfriend.aspx

Heffernan, Troy. 2021. „Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching". In: *Assessment & Evaluation in Higher Education*. DOI: 10.1080/02602938.2021.1888075

Heilman, Madeline E. und Michelle C. Haynes. 2005. „No credit where credit is due: attributional rationalization of women's success in male-female teams". In: *Journal of Applied Psychology* 90, 5: 905-916.

Heilman, Madeline E. und Tyler G. Okimoto. 2007. „Why are women penalized for success at male tasks? The implied communality deficit". In: *Journal of Applied Philosophy* 92, 1: 81-92.

Heilman, Madeline E.; Wallen, Aaron S.; Fuchs, Daniella und Melinda M. Temkins. 2004. „Penalties for success: reactions to women who succeed at male gender-typed tasks". In: *Journal of Applied Psychology* 89, 3: 416-27.

Himelein, Melissa J. 2017. „Pitfalls of using student comments in the evaluation of faculty". *Academic Briefing*. Letzer Zugriff 16. Februar 2022. https://www.academicbriefing.com/human-resources/faculty-evaluation/pitfalls-of-using-student-comments-evaluation-of-faculty/

Hornstein, Henry A. 2017. „Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance". In: *Cogent Education* 6, 1: https://doi.org/10.1080/2331186X.2017.1304016.

Joye, Shauna W. und Janie H. Wilson. 2015. „Professor age and gender affect student perceptions and grades". In: *Journal of the Scholarship of Teaching and Learning* 15, 4:126-138.

Kaatz, Anna; Gutierrez, Belinda und Molly Carnes. 2014. „Threats to objectivity in peer review: the case of gender". In: *Trends in Pharmacological Sciences* 35, 8: 371-373.

Klonschinski, Andrea. 2018. „#MeToo und Frauen in der akademischen Philosophie: der perfekte Sturm“. Philosophie-Blog *praefaktisch*. Letzter Zugriff 15. Februar 2022. https://www.praefaktisch.de/metoo/metoo-und-frauen-in-der-akademischen-philosophie-der-perfekte-sturm/

Klonschinski, Andrea. 2020. „Frauen in der Deutschen Akademischen Philosophie - Eine Bestandsaufnahme“. In: *Zeitschrift für Philosophische Forschung* 74, 4: 593-616.

Krawczyk, Michael. 2017. „Are all researchers male? Gender misattributions in citations“. In: *Scientometrics* 110: 1397-1402.

Krawczyk, Michael und Magdalena Smyk. 2016. „Author's gender affects rating of academic articles: evidence from an incentivized, deception-free laboratory experiment“. In: *European Economic Review* 90: 326-335.

Kreitzer, Rebecca J. und Jennie Sweet-Cushman. 2021. „Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform“. In: *Journal of Academic Ethics 20:* 73-84. https://doi.org/10.1007/s10805-021-09400-w.

Leslie, Sarah-Jane; Cimpian, Andrei; Meyer, Meredith; Freeland, Edward. 2015. „Expectations of brilliance underlie gender distribution across academic disciplines“. In: *Science* 347, 6219: 262-265.

Linka, Albert N.; Swanna, Christopher A. und Barry Bozeman. 2008. „A time allocation study of university faculty“. In: *Economics of Education Review* 27: 363–374.

MacNell, Lillian; Driscoll, Adam und Andrea N. Hunt. 2015. „What's in a name: exposing gender bias in student ratings of teaching“. In: *Innovative Higher Education*, 40: 291-303.

Manne, Kate. 2018. *Down Girl: The Logic of Misogyny*. Oxford u. a.: Oxford University Press.

Mengel, Friederike; Sauermann, Jan und Ulf Zölitz. 2019. „Gender bias in teaching evaluations“. In: *Journal of the European Economic Association* 17, 2: 535–566.

Miles, Patti und Deanna House. 2015. „The tail wagging the dog: an overdue examination of student teaching evaluations“. In: *International Journal of Higher Education* 4, 2: 116-126.

Misra, Joya; Hickes Lundquist, Jennifer; Holmes, Elissa und Stephanie Agiomavritis. 2011. „The ivory ceiling of service work“. American Association of University Professors, Januar-Februar. Letzter Zugriff am 15. Februar 2022. https://www.aaup.org/article/ivory-ceiling-service-work#.W5jTRxEyW70

Mitchell, Kristina M. W. und Jonathan Martin. 2018. „Gender bias in student evaluation“. In: *Political Science and Politics* 51, 3: 648-652.

Moss-Racusin, Corinne A.; Dovidio, John F.; Brescoll, Victoria L.; Graham, Mark J. und Jo Handelsman. 2012. „Science faculty's subtle gender biases favor male students“. In: *PNAS* 109, 41: 16474-16479.

Murray, Dakota; Boothby, Clara; Zhao, Huimeng; Minik, Vanessa; Bérubé, Nicolas; Larivière, Vincent und Cassidy R. Sugimoto. 2020. Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS ONE* 15, 6: e0233515. https://doi.org/10.1371/journal.pone.0233515

Peterson, David A. M., Biederman, Lori A., Andersen, David, Ditonto, Tessa M. und Kevin Roe. 2019. „Mitigating gender bias in student evaluations of teaching“. *PloS ONE* 14, 5: e0216241. https://doi.org/10.1371/journal.pone.0216241

Rivera, Lauren A. und András Tilcsik. 2019. „Scaling down inequality: rating scales, gender bias, and the architecture of evaluation“. In: *American Sociological Review* 84, 2: 248-274.

Rosen, Andrew S. 2017. „Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data“. In: *Assessment & Evaluation in Higher Education* 43, 1: 1-14.

Sarson, Heather. 2017. „Gender differences in recognition for groups“. Working Paper, Harvard University. https://scholar.harvard.edu/files/sarsons/files/full_v6.pdf.

Saul, Jennifer. 2013. „Implicit bias, stereotype threat, and women in philosophy". In: Hutchison, Katrina und Fiona Jenkins (Hg.): *Women in Philosophy. What Needs to Change?* 39-60. Oxford u. a.: Oxford University Press.

Sprague, Joey und Kelley Massoni. 2005. „Student evaluations and gendered expectations: what we don't know can't hurt us". In: *Sex Roles* 53, 11/12: 779-793.

Stark, Philip B. und Richard Freishtat. 2014. „An evaluation of course evaluations". *ScienceOpen Research.* DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

Staufenbiel, Thomas; Seppelfricke, Thomas und Judith Rickers. 2015. „Prädikatoren studentischer Lehrveranstaltungsevaluationen. Eine Mehrebenenanalyse". In: *Diagnostica* 62, 1: 44-59.

Storage, Daniel; Horne, Zachary; Chimpian, Andrei und Sarah-Jane Leslie. 2016. „The frequency of „brillant" and „genius" in teaching evaluations predicts the representation of women and African Americans across fields". In: *PLoS ONE* 11, 3: e0150194. https://doi.org/10.1371/journal.pone.0150194

Trix, Frances und Carolyn Psenka. 2003. „Exploring the color of glass: letters of recommendation for female and male medical faculty". In: *Discourse and Society* 14: 191-220.

Tucker, Beatrice. 2014. „Student evaluation surveys: anonymous comments that offend or are unprofessional". In: *Higher Education* 68: 347-358.

Uttl, Bob, White, Carmela A. und Daniela Wong Gonzalez. 2017. „Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related". In: *Studies in Educational Evaluation* 54: 22-42.

Valian, Virginia. 1999. *Why so Slow? The Advancement of Women.* Cambridge und London: MIT Press.

Valian, Virginia. 2005. „Beyond gender schemas: improving the advancement of women in academia". In: *Hypatia* 20, 3: 198-213.

Wagner, Natascha; Rieger, Matthias und Katherine Voorvelt. 2016. „Gender, ethnicity and teaching evaluations: evidence from mixed teams". In: *Economics of Education Review* 54: 79-94.

Wolbring, Tobias. 2010. „Attraktivität, Geschlecht und Lehrveranstaltungsevaluation. Eine Replikationsstudie zu den Befunden von Hamermesh und Parker (2005) und Klein und Rosar (2006) mit Hilfe von Individualdaten". In: *Zeitschrift für Evaluation* 9, 1: 29-48.

Wolfers, Justin. 2016. „When team-work doesn't work for women". *New York Times.* Januar. Letzter Zugriff 16. Februar 2022. https://www.nytimes.com/2016/01/10/upshot/when-teamwork-doesnt-work-for-women.html

Kommentar

Prof. Dr. Liudvika Leišytė[1]

[1] TU Dortmund; liudvika.leisyte@tu-dortmund.de

Viele Studien haben gezeigt, dass die Bewertung von Forschung, Lehre und Transfer in der Akademie gegendert ist und somit systematische Hindernisse für Frauen beim Zugang zu hohen Positionen in der Wissenschaft schafft (Savigny 2014; Jappelli et al. 2015; Rivera & Tilcsik 2019; Leišytė/Peksen 2019). Dr. Klonschinskis Artikel, der auf einer gründlichen Literaturrecherche zur Evaluierung der Hochschulehre basiert, belegt eindeutig den Gender Bias bei der Bewertung von Lehrveranstaltungen in verschiedenen Ländern und Disziplinen. Es ist ein sehr begrüßenswerter Beitrag, der einmal mehr die Allgegenwärtigkeit von Gender Bias in Evaluierungen im akademischen Bereich hervorhebt. Der Artikel stellt das üblicherweise „normalisierte" Muster der Lehrevaluierung in Frage, das von den Universitäten oft aufgrund der Anforderungen der Programm- oder institutionellen Akkreditierung zur Überwachung der Lehrleistung und nur selten mit dem Ziel der Verbesserung der Qualitätskultur eingesetzt wird. Wie Dr. Klonschinski zu Recht bemerkt hat, sind Lehrevaluationen vielfach kritisiert worden, weil sie die Qualität von Lehre und Lernen nicht messen. Trotz all der langjährigen Kritik ist die Lehrevaluation nach wie vor das wichtigste Instrument zur „Messung" der Lehrleistung des akademischen Personals in den Hochschulsystemen weltweit. Darüber hinaus haben trotz der zahlreichen Studien, die vor „Zufriedenheituntersuchungen" über Studierende warnen, haben diese Art von Studien einen Einfluss darauf, welche Lehrinhalte und Lehrmethoden verwendet werden um sicherzustellen, dass an den Universitäten, die „beste studentische Erfahrung" gewährleistet werden kann, wie im Vereinigten Königreich oder in den Niederlanden. Leider sind bei der Konzeption von Evaluierungsinstrumenten im Hochschulbereich, sei es in der Lehre oder in der Forschung, oftmals nicht die Vielzahl an möglichen Bias, wie z. B. Gender Bias, der in der Entwicklung von Messinstrumenten eingebettet wird, bewusst. Daher ist die Forderung nach einer Sensibilisierung und einer politischen Debatte über Gender Bias in der Lehr- und Forschungsevaluation zeitgemäß und höchst relevant.

Die Autorin stellt die Ergebnisse einer Reihe von Studien aus den USA, Frankreich und den Niederlanden vor, von denen viele, die auf experimentellen Designs basieren, auf signifikante geschlechtsspezifische Unterschiede in den Ergebnissen der Lehrevaluation hinweisen. Die vorgelegte empirische Grundlage ist umfangreich und überzeugend. Besonders interessant finde ich, dass Statusgruppen eine Rolle dabei spielen, wie signifikant der Geschlechterunterschied in der Lehrevaluation ist. Hier sollten die übergreifenden Effekte von Geschlecht und Alter hervorgehoben werden, nicht nur von Geschlecht und Status. Insgesamt würde ich auch argumentieren, dass es sehr wichtig ist, zu unterscheiden, inwieweit der Gender Bias nicht nur von der Karrierestufe, sondern auch von der Disziplin, verschiedenen Hintergrundkategorien der Studierenden (nicht nur vom Geschlecht) und dem kulturell-ökonomischen Kontext abhängt. Unsere Studie zur Bewertung von Online-Lernumgebungen in Schweden und Deutschland hat beispielsweise gezeigt, dass sowohl das Geschlecht als auch das Land einen Unterschied in der Wahrnehmung der Qualität von Online-Lernumgebungen ausmachen (Waheed/Leisyte 2021). Auch die Bewertung des Unterrichts ist unterschiedlich, je nachdem, ob er online stattfindet oder nicht, und die Art und Weise, wie der Unterricht in einem bestimmten System organisiert ist macht

ebenfalls einen großen Unterschied. Darüber hinaus spielen Stereotypen in verschiedenen Kulturen eine Rolle, und dies muss berücksichtigt werden. Die meisten Studien konzentrieren sich auf einen bestimmten kulturellen oder disziplinären Kontext, so dass mehr vergleichende Arbeiten über die verschiedenen Kulturen, die verschiedenen Arten von Einrichtungen und Disziplinen hinweg dringend notwendig sind.

Eine weitere wichtige Erkenntnis ist die unterschiedliche Bewertung von männlichen und weiblichen Wissenschaftlern, wobei männliche Wissenschaftler als „Entertainer" und weibliche Wissenschaftlerinnen als „Nurturer" eingestuft werden. Die mangelnde Wertschätzung der Lehre von Frauen durch Studierende ist ein wichtiges Ergebnis, das Wissenschaftlerinnen davon abhalten könnte, eine wissenschaftliche Laufbahn einzuschlagen. In diesem Zusammenhang halte ich es auch für besonders wichtig, Studien einzubeziehen, die untersuchen, wie Entscheidungsträger Evaluierungen nutzen, um Einstellungs- und Beförderungsentscheidungen zu treffen. Die Studien zum „Peer Review Bias" (Lamont 2009) zeigen, wie wichtig es ist, auch die Verzerrungen bei der Verwendung von Evaluationsergebnissen für Finanzierungs-, Einstellungs- oder Beförderungsentscheidungen zu entschlüsseln. Studien zu diesem Bereich im deutschen Hochschulsystem sind eher selten.

Literatur

Jappelli, Tullio, Nappi, Carmela A. und Torrini, Roberto. 2015. Research Quality and Gender Gap in Research Assessment. Arbeitspapier Nr. 418. Neaple: Universität Neapel Centre for Studies in Economics and Finance. Letzter Zugriff am 01.08. 2022. http://www.csef.it/WP/wp418.pdf

Lamont, Michèle. 2009. *How Professors think?* Cambridge: Harvard University Press.

Leišytė, Liudvika und Peksen, Sude. 2020. „Nationale Evaluationssysteme für Forschung in Hochschulen – Gender Bias im europäischen Vergleich." In: Isabell M. Welpe; Jutta Stumpf-Wollersheim; Nicholas Folger; Manfred Prenzel (Hg.): *Leistungsbewertung in wissenschaftlichen Institutionen und Universitäten: Eine mehrdimensionale Perspektive*: 13–41. Berlin: De Gruyter.

Neave, Guy. 2012. „The Evaluative State: A Formative Concept and an Overview." In: Guy Neave (Hg.), *The Evaluative State, Institutional Autonomy and Re-engineering Higher Education in Western Europe.* 36-46. London: Palgrave Macmillan.

Rivera, Lauren A. und Tilcsik, András. 2019. „Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." In: *American Sociological Review*, 84, 2:248–274

Savigny, Heather. 2014. „Women, know your limits: Cultural sexism in academia." In: *Gender and Education*, 26, 7:794–809.

Waheed, Mehwish und Leišytė, Liudvika. 2021. „German and Swedish students going digital: Do gender and interaction matter in quality evaluation of digital learning systems?" In: *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1965626