erschienen als: Reinhart, M. (2023) 'Wertvolle Forschung: Die Konstruktion, Produktion, Bewertung und Sicherung wissenschaftlicher Qualität', in D. Kaldewey (Hrsg.) Wissenschaftsforschung. De Gruyter Oldenbourg, S. 203–220. https://doi.org/10.1515/9783110713800-010.

Wertvolle Forschung – Die Konstruktion, Produktion, Bewertung und Sicherung wissenschaftlicher Qualität

Martin Reinhart
Robert K. Merton Zentrum für Wissenschaftsforschung
Humboldt-Universität zu Berlin
martin.reinhart@hu-berlin.de

Einleitung

Die Qualität von Forschung wird in der Wissenschaft fortlaufend bewertet, wobei dieses Bewerten gleichermaßen auffällig sichtbar ist, aber auch gezielt unsichtbar gehalten wird. Für herausragende Forschung werden Preise verliehen, wie bspw. der Nobelpreis, die über die Wissenschaft hinaus für An- und Aufsehen sorgen. Universitäten werden bezüglich ihrer Qualität in Hochschulrankings bewertet und verglichen, worüber massenmedial nicht nur prominent berichtet wird, sondern was die Erfolgreichen auch dazu veranlasst, ihre Position als Ausweis ihrer Leistung und als Zeichen ihrer Reputation öffentlich herauszustellen. Augenfällig ist auch, dass Forscher*innen ihre Lebensläufe mit umfangreichem Leistungsausweis an Publikationen, Forschungsprojekten oder Vorträgen auf Webseiten präsentieren, als ob sie im Rahmen eines permanenten Stellenbewerbungsverfahrens die Bewertung ihrer individuellen Qualitäten einfordern würden. Sogar schlechte Forschungsqualität und deren Bewertung sind auffällig, ja gar spektakulär, sichtbar, wenn Irrtümer oder Fälle von wissenschaftlichem Fehlverhalten publik und als massenmediale Skandale inszeniert werden.

Während die Ergebnisse von Bewertungsprozessen oft deutlich sichtbar sind, so sind es die Prozesse, die zu diesen Ergebnissen führen, meist nicht. Kommissionen, die Preise vergeben, Doktortitel gewähren, Professuren besetzen oder Mittel für Forschungsprojekte zuweisen, tagen vornehmlich unter Ausschluss der Öffentlichkeit. Dabei stützen sie sich oft auf Gutachten, die von Expert*innen vertraulich verfasst sind und manchmal auch den darin Bewerteten nicht zur Kenntnis gegeben werden. Auch Gutachten zur Bewertung von eingereichten Manuskripten bei Zeitschriften oder Verlagen bleiben meist vertraulich, so dass Gutachten als Genre der schriftlichen Bewertung wissenschaftlicher Qualität zwar als eine der häufigsten Textsorten in der Wissenschaft überhaupt gelten können, die aber gleichzeitig nur einem sehr kleinen Publikum zugänglich sind.

Neben dieser Vielfalt ist es die Häufigkeit, mit der begutachtet, bewertet, evaluiert wird, die eine Auffälligkeit des Wissenschaftssystems darstellt. Robert K. Merton hat dafür den Begriff des organisierten Skeptizismus als eine der fundamentalen normativen Orientierungen in der Wissenschaft geprägt (Merton 1973 [1942]).

Box X.1: Mertons Ethos der Wissenschaft

Was sind die moralischen Imperative, denen sich Forschende verpflichtet fühlen? Merton stellt sich diese Frage 1942 unter dem Eindruck der Bedrohung wissenschaftlicher Autonomie durch Krieg und Diktatur. Er identifiziert vier institutionelle Imperative, die zusammen als moralischer Kern eines wissenschaftlichen Ethos geglaubt werden: **Universalismus** fordert auf, wissenschaftliche Aussage nach unpersönlichen Kriterien und damit unter Absehung von Kultur, Nationalität, Geschlecht, etc. zu bewerten. **Kommunalismus** fordert auf, Wissenschaft als gemeinschaftliches Unterfangen zu verstehen und sich auf eine offene Form der Kommunikation frei von Geheimhaltung und kapitalistischen Verwertungsinteressen zu verpflichten. **Desinteressiertheit** fordert auf, wissenschaftliche Aussagen rigoros auf Wahrheit und mögliche Einflussnahme zu prüfen. **Organisierter Skeptizimus** schliesslich fordert auf, sowohl die Forschung wie die Institutionen der Wissenschaft so zu organisieren, dass Universalismus, Kommunalismus und Desinteressiertheit dauerhaft Genüge getan wird. Forschende halten diese Normen nicht nur deshalb für verbindlich, weil sie der Erweiterung bestätigten Wissens zuträglich sind, sondern weil sie für gut und richtig gehalten werden, so Merton.

Auffällig sind aber auch die Sichtbarkeitsverhältnisse rund um das Bewerten von wissenschaftlicher Qualität, die nicht einfach auf maximale Transparenz zielen, sondern das Resultat einer historisch gewachsenen Struktur der Qualitätssicherung

darstellen. Deren Funktion ist nicht nur die Sicherung der Qualität des produzierten Wissens, indem Fehler korrigiert, methodische Standards aufrechterhalten oder originelle Beiträge als solche erkannt werden. Sie weist auch Reputation zu, sie fördert Kollegialität im Wettbewerb um diese Reputation und sie schafft einen professionellen Raum, in dem Wissensproduktion in relativer Autonomie zu gesellschaftlichen Ansprüchen bspw. aus der Politik, der Wirtschaft oder den Massenmedien stattfinden kann. Fragen danach, wie Qualität in der Wissenschaft bewertet wird, sind damit immer auch Fragen danach, wie die Wissenschaft gesteuert und regiert wird.

Phänomene wissenschaftlichen Bewertens

Auf den ersten Blick scheint das Bewerten von wissenschaftlicher Qualität ein sehr heterogenes Phänomen zu sein (Schendzielorz & Reinhart 2020). Wie oben angedeutet können darunter so unterschiedliche Dinge wie die öffentliche Vergabe von Preisen für herausragende Leistungen oder die streng vertrauliche Untersuchung eines Verdachts auf wissenschaftliches Fehlverhalten fallen. Während diese beiden Formen des Bewertens selten stattfinden, so gehören andere Formen zum wissenschaftlichen Alltag. Niederschwellige und alltägliche Formen des Bewertens finden sich in jeder Forschungsgruppe, in der Kooperation und kollegialer Austausch aber auch Konkurrenz und Streit stattfinden. In Flurgesprächen, beim Mittagessen oder beim unmittelbar gemeinsam Forschen drehen sich Gespräche immer wieder darum, ob eine benutzte Methode verbessert werden kann, ob die bestmöglichen Forschungsmaterialien verwendet werden, ob jemand von neuen und besseren Experimenten oder Argumenten weiß, bis hin zu hitzigen Debatten, ob ein bestimmtes Vorgehen noch dem Stand des Fachs entspricht oder überhaupt in der Lage ist, einen relevanten Beitrag zur anvisierten Forschungsfrage zu leisten. So beiläufig diese Formen scheinen mögen, so kommt in ihnen doch zum Ausdruck, dass zum wissenschaftlichen Alltag und für einen wissenschaftlichen Habitus eine kritische Grundstimmung gehört, die im Extremfall jeden Arbeitsschritt einer erneuten Prüfung unterziehen kann. Ein taxierender, auf Qualität gerichteter Blick gehört somit zum geteilten Grundverständnis von Wissenschaftler*innen. Das heißt nicht, dass im Forschungsalltag keine Routinen und Konventionen zu finden sind, denen selbstverständlich einfach gefolgt wird. Es heißt vielmehr, dass diese Routinen und Konventionen eine erhöhte Chance haben, problematisiert zu werden, falls sich dieser taxierend bewertende Blick aus Qualitätsgründen auf sie richtet. Ein Beispiel hierfür ist die aktuelle Kritik an

statistischen Standardmethoden, die im letzten Kapitel diskutiert werden wird.

Weniger beiläufige Formen des Bewertens finden sich überall da, wo Forschungsvorhaben und Forschungsergebnisse explizit zur Diskussion gestellt werden. Der wissenschaftliche Alltag ist durchsetzt von Besprechungen in Forschungsgruppen, Workshops zum Austausch mit Fachkolleg*innen, Kolloquien zu laufenden Promotionsvorhaben, bis hin zu den regelmäßigen Konferenzen von nationalen und internationalen Fachgesellschaften, an denen Vorträge gehalten werden. Als Selbstverständlichkeit gehört zu diesen Vorträgen, dass an sie eine Diskussion anschließt, in der Verständnisfragen gestellt, kritische Diskussionsbedarfe angemerkt und manchmal auch Bewunderung oder Verachtung zum Ausdruck gebracht werden können. Schließt an einen Vortrag keine Diskussion an, in der dieser taxierende bewertende Blick zum Ausdruck kommt, stellt sich im wissenschaftlichen Feld schnell die Frage, ob es sich dabei denn überhaupt um einen wissenschaftlichen Vortrag gehandelt habe. Im Rahmen von wissenschaftlichen Konferenzen zeigt sich das schon in der Vorbereitung, indem Bewertung und Kritik stärker organisiert und formalisiert werden. Forschende bewerben sich darum, an einer Konferenz vortragen zu dürfen, indem sie ein *Abstract* oder ein Manuskript einreichen, das von Fachkolleg*innen begutachtet wird.

Solche Begutachtungsverfahren, in denen ein Ergebnis wissenschaftlicher Arbeit eingereicht und von Fachkolleg*innen bewertet wird, werden allgemein als *Peer Review* bezeichnet (Neidhardt 2016). Bewertungsgegenstand können dabei nicht nur *Abstracts* für Konferenzvorträge sein, sondern auch Zeitschriftenartikel, Bücher, Anträge für Forschungsprojekte oder Stellenbewerbungen. *Peer Review*-Verfahren kommen deshalb bei Konferenzen, bei Zeitschriften, bei Verlagen, in der Forschungsförderung und bei Stellenbesetzungen zum Einsatz. Analoge Verfahren werden in der Wissenschaft auch zur institutionellen Evaluation verwendet, also bspw. wenn die Arbeit eines ganzen Forschungsinstituts, einer großen Forschungskooperation mit zahlreichen Partnerinstitutionen oder die Ergebnisse eines ganzen Programms zur Forschungsförderung bewertet werden sollen. Was dabei als *Peer Review* bezeichnet wird, kann sehr vielfältig sein. Gemeinsam ist allen Verfahren, dass ein Bewertungsobjekt vorgelegt werden muss („Postulat“), dass eine Bewertung auf der Basis von fachbezogener Expertise stattfinden muss und dass eine Entscheidung über Publikation, Förderung, Einstellung, etc. anschließt (Schendzielorz & Reinhart 2020: 115ff.). Diese drei Verfahrensschritte können aber sehr unterschiedlich ausgeprägt sein und es können auch zusätzliche Schritte hinzukommen. Die Begutachtung kann durch eine feste Gruppe erfolgen oder durch immer wieder neu ausgewählte Fachgutachten-

de. Bei der Deutschen Forschungsgemeinschaft (DFG), die staatliche Mittel für Grundlagenforschung vergibt, gibt es sogar beides in Kombination: gewählte Fachvertreter*innen, die als Fachkollegien organisiert begutachten, die aber bei Bedarf externe Gutachtende hinzuziehen.

Box X.2 Die Deutsche Forschungsgemeinschaft (DFG)

In Deutschland wird Forschung durch öffentliche Mittel (Bund und Länder) primär auf zwei Wegen finanziert: Einerseits als Grundhaushalt für Universitäten und Ausseruniversitäre Forschungseinrichtungen (bspw. der Max-Planck-Gesellschaft) andererseits als wettbewerblich vergebene Drittmittel über die DFG. Die DFG ist als eingetragener Verein organisiert und beschreibt sich als „Selbstverwaltungsorganisation der Wissenschaft“. Anträge von Forschenden werden anhand disziplinärer Zugehörigkeit in 48 Fachkollegien begutachtet und zur Bewilligung oder Ablehnung vorgeschlagen. Dem Anspruch auf wissenschaftliche Selbstverwaltung wird dadurch Rechnung getragen, dass die Mitglieder der Fachkollegien durch die Forschenden in den jeweiligen Fächern für vier Jahre gewählt werden. Zum Hauptausschuss, der über die Vorschläge der Fachkollegien definitiv entscheidet, gehören neben Wissenschaftler*innen aber auch Vertreter*innen von Bund und Ländern. Die DFG hat nicht nur Förderprogramme für individuelle Forschungsprojekte sondern auch für Forschungskollaborationen (bspw. Sonderforschungsbereiche) oder für individuelle Karrieren (bspw. Emmy Noether-Programm). Die DFG ist ein gutes Beispiel für das, was weiter unten als Grenzorganisation beschrieben wird.

Die Begutachtung kann durch individuelles Lesen des Postulats und Verfassen eines schriftlichen Gutachtens erfolgen, aber auch durch Diskussion einer Gruppe von Gutachtenden (*Panel*), die vor Ort stattfindet. Eine solche *Panel*-Begutachtung enthält manchmal auch das Element, dass die Bewerteten ihr Postulat vor Ort präsentieren und mit den Gutachtenden diskutieren. Extrem aufwändige Formen der Begutachtung können zehn oder mehr solcher Verfahrensschritte beinhalten und müssen aufwändig organisiert werden: Die Begutachtung in der Exzellenzstrategie für deutsche Universitäten (Möller et al. 2012) oder die Vergabe von Personenförderung durch das *European Research Council* (ERC) wären Beispiele dafür (Reinhart & Schendzielorz 2021a).

Was die konkreten Eigenschaften wissenschaftlicher Arbeit sind, die das Attribut „Qualität“ oder „qualitativ hochwertig“ tragen können, ist schwer zu bestimmen (Dahler-Larsen 2019). Häufig genannte Aspekte sind bspw. Originalität, methodische Strenge oder Widerspruchsfreiheit. Im wissenschaftlichen Alltag werden Qualitätskriterien kaum explizit definiert und oft mit „I

know it, when I see it“ ins Auge der Betrachtenden verlagert. Dazu gehört auch, dass meist davon ausgegangen wird, dass sich Qualitätsvorstellungen zwischen den Disziplinen stark unterscheiden, aber trotzdem unterstellt wird, dass es basale Qualitäten gibt, die wissenschaftliche und nichtwissenschaftliche Arbeit abgrenzen können. In der Folge sind Versuche der Wissenschaftsforschung, Qualitätsvorstellungen zu rekonstruieren, nur begrenzt ergiebig (Guetzkow et al. 2004, Hug & Aeschbach 2020). Diese Unterbestimmtheit führt in der Bewertungspraxis oft dazu, dass einfacher bestimmbare Kriterien herangezogen werden, die man allenfalls als sekundäre Qualitäten bezeichnen könnte. So werden die Anzahl Zitate als Indikatoren für die Qualität von Publikationen genutzt oder eingeworbene Drittmittelsummen für die Qualität von Forschungsprojekten. Wie sich unten zeigen wird, macht es deshalb Sinn, davon auszugehen, dass es die Bewertungsverfahren selbst sind, die wissenschaftliche Qualität hervorbringen, womit die Performativität und Reflexivität des Bewertens in Rechnung gestellt wird.

Bewerten als gesellschafts- und sozialtheoretisches Phänomen

Bewertungsprozesse sind grundlegend dafür wie wir als Individuen und Kollektive die Welt wahrnehmen und gestalten (Krüger & Reinhart 2016). „Man macht sich selten klar, dass unser ganzes Leben [...] in Wertgefühlen und Wertabwägungen verläuft und überhaupt nur dadurch Sinn und Bedeutung bekommt, dass die mechanisch abrollenden Elemente der Wirklichkeit über ihren Sachgehalt hinaus unendlich mannigfaltige Maße und Arten von Wert für uns besitzen“ (Simmel 2008 [1900]: 25). Denkt man in dieser Allgemeinheit über das Phänomen nach, so lässt sich das Bewerten in zwei Momente zerlegen: Zuschreiben („Wertgefühle“) und Vergleichen („Wertabwägungen“). Der Wert eines Objekts besteht damit nicht einfach in einer intrinsischen Eigenschaft, bspw. der Schönheit eines Gemäldes oder der Seltenheit eines Rohstoffes, sondern darin, dass dem Objekt zugeschrieben wird, überhaupt wertvoll genug zu sein, um bewertet zu werden, und es sich dann im Vergleich zu anderen ähnlichen Objekten einordnen lässt. In der Praxis lassen sich diese Momente jeweils nur schwer auseinander halten: Eine Rangliste von vielzitierten Publikationen impliziert bspw. gleichzeitig, dass Publikationen in der Wissenschaft an sich etwas Wertvolles sind und dass sie sich anhand von Zitationen miteinander vergleichen lassen. Ein derartiges Zusammenspiel des Zuschreibens und Vergleichens von Wert bildet die Grundlage dafür, was gesellschaftlich als sinnvoll oder bedeutsam angesehen werden kann (Krüger und Reinhart 2017).

In diesem allgemeinen Sinn ist das Bewerten ein fast unsichtbares Phänomen, weil soziale Akteure in ihrem Tun nicht anders können, als laufend Wertzuschreibungen und -abwägungen vorzunehmen. Gerade weil dies laufend und implizit geschieht, treten immer wieder Störungen auf, wenn unterschiedliche Bewertungen aufeinander treffen. Eine Universität ist bspw. nicht mit der Rangierung im neuesten Hochschulranking einverstanden und hinterfragt damit nicht nur das Ergebnis sondern auch die Methode der vorgenommenen Bewertung. Über solche Störungen und Widersprüche kommt ein gesellschaftlicher Prozess in Gang, der die Bewertungsprozesse selbst problematisiert und auf deren Explizierung oder gar Formalisierung drängt. Das Resultat davon sind Bewertungs*verfahren* oder wie der französische Neopragmatismus sagen würde: „Tests" (Potthast 2017). Verfahren oder Tests explizieren das Bewerten nicht nur in sichtbarer Weise, sie machen es auch kritisierbar und damit gestaltbar. Beispiele hierfür finden sich in allen gesellschaftlichen Bereichen, aber in Wissenschaft und Bildung sind sie besonders prägnant (Schulnoten, Abitur, Promotion, Berufsabschlüsse, etc.). Die Sichtbarkeit und Erwartbarkeit dieser Tests führt dazu, dass diejenigen, die sich ihnen unterziehen müssen oder wollen, sich darauf vorbereiten. Durch das Lernen aufs Abitur oder durch das Schreiben einer Dissertation zur Promotion produzieren die zu Bewertenden schon im Vorgriff auf das eigentliche Bewertungsverfahren, die eingeforderten Qualitäten. Für den Fall der Hochschulrankings wird dies von Sauder & Espeland (2009) eindrücklich dargestellt. In diesem Sinne messen Bewertungsverfahren nicht einfach die Qualität von etwas, sondern bringen diese überhaupt erst hervor (Schendzielorz & Reinhart 2020: 113-115).

Damit lässt sich sagen, dass Bewertungsverfahren performativ und reflexiv sind. Performativ sind sie in dem Sinne, dass Akteure sich ihnen unterwerfen (müssen) und damit erzeugen sie, was die Verfahren zu bestimmen versuchen. Reflexiv sind sie, indem sie den Akteuren ermöglichen durch Kritik zur Umgestaltung der Verfahren oder durch Boykott zu deren Abschaffung beizutragen.[1] Die Wissenschaftsforschung adressiert diese Themen, indem sie einerseits nach den gesellschaftlichen Bedingungen und Folgen der Performativität und Reflexivität von Bewertungsverfahren fragt. Power spricht in diesem Zusammenhang von einer „audit society" (Power 1997), Dahler-Larsen von einer „evaluation society" (Dahler-Larsen 2012) und zunehmend tauchen jetzt auch Diagnosen eines „Plattformkapitalismus" (Mirowski 2018) auf. Das Phänomen bleibt entsprechend nicht auf die Wissenschaft beschränkt, sondern betrifft Gesell-

1 Die hier zugrunde gelegte Unterscheidung in "exit, voice, and loyalty" stammt von Hirschman (1970). Eine weiterführende theoretische Auseinandersetzung zur Reflexivität öffentlicher Kritik findet sich bei Boltanski & Thévenot (1999).

schaft als Ganzes. Herausgebildet hat sich auch eine spezialisierte Literatur, die stärker am Funktionieren einzelner dieser Bewertungsverfahren interessiert ist. Lamont bezeichnet diese als „comparative sociology of valuation and evaluation“ (Lamont 2012), die mit „Valuation Studies“ seit 2013 auch eine eigene Zeitschrift kennt (https://valuationstudies.liu.se/).

Peer Review als spezifisches Bewertungsformat der Wissenschaft

Augenfälligstes Bewertungsverfahren in der Wissenschaft ist das *Peer Review*. Dieses gibt es in unzähligen Formen, so dass sich zurecht die Frage stellen lässt, ob diese alle unter einem Begriff subsumiert werden können. Die Bewertungsobjekte können sehr unterschiedlich sein: Manuskripte, Projektanträge, Lebensläufe oder institutionelle Selbstberichte. In die Begutachtung können nur wenige Expert*innen involviert sein, die aus der Ferne ein Gutachten verfassen, oder eine größere Gruppe, die vor Ort als *Panel* tagt. Die notwendige Expertise zur Bewertung kann sehr eng und fachbezogen, aber auch nur sehr allgemein wissenschaftsbezogen sein. Aufgrund dieser Heterogenität ist eine Konvention in der Wissenschaft, v.a. die Begutachtung bei Zeitschriften und bei der Forschungsförderung als *Peer Review* zu bezeichnen, während institutionelle Evaluationen und Stellenbesetzungsverfahren oft ausgeklammert werden. Im Gegensatz dazu soll hier aus drei Gründen an einem umfassenderen *Peer Review*-Begriff festgehalten werden: 1. Theoretisch lässt sich trotz der Heterogenität argumentieren, dass all diesen Verfahren ähnliche Steuerungsfunktionen zukommen, die die relative Autonomie der Wissenschaft begründen. Braun spricht in diesem Zusammenhang vom „Dualismus von Regulierungsanspruch und korporatistischer Selbstverwaltung“ (Braun 1997: 100, siehe auch Neidhardt 2016). 2. Praktisch lässt sich zeigen, dass sich die Verfahren als Ganzes zwar stark unterscheiden, dass sie aber jeweils aus ähnlichen Verfahrenselementen zusammengesetzt sind (Schendzielorz und Reinhart 2020). 3. Methodisch besteht das Problem, dass die Wissenschaftsforschung zum Peer Review vor allem aus Fallstudien zu einzelnen Verfahren besteht, so dass ein Mangel an vergleichenden Arbeiten einen fragmentierten Forschungsstand zurück lässt (Hirschauer 2004). Es ist vor allem das Peer Review bei Zeitschriften und in geringerem Umfang in der Forschungsförderung, zu dem zahlreiche empirische Forschungsarbeiten vorliegen.

Der Stand der Forschung zum *Peer Review* ist ungewöhnlich, weil vergleichsweise wenige empirische Studien vorliegen, die eindeutig in der Literatur der Wissenschaftsforschung verankert sind (Reinhart 2012). Viele Arbeiten stammen von (ehe-

maligen) Verfahrensbeteiligten (bspw. Zeitschrifteneditoren), die dadurch einen privilegierten Datenzugang haben und mit primär praxisorientierten Fragestellungen arbeiten (Wie lässt sich unser Verfahren verbessern?). Auch sind viele Arbeiten einem Defizitmodell verpflichtet; gehen also von einem spezifischen Mangel des Peer Review aus und suchen diesen zu bestätigen resp. Vorschläge zu dessen Behebung zu machen (Reinhart & Schendzielorz 2021). Trotzdem lassen sich aus der Summe dieser Arbeiten einige Schlussfolgerungen ziehen (Guthrie et al. 2017, Neidhardt 2016): Ob Peer Review-Verfahren valide sind, also gute und schlechte Qualität unterscheiden können, lässt sich direkt nicht beantworten, u.a. weil Verfahren wie oben beschrieben performativ und reflexiv sind. Zeigen lässt sich aber, dass es eher die Ausnahme als die Regel ist, dass sich Gutachtende in ihrem Urteil einig sind (geringe Reliabilität). Gut bestätigt ist auch, dass sich die Auswahlentscheidungen erheblich ändern, wenn andere aber gleichermaßen qualifizierte Gutachtende zum Einsatz kommen. Valide scheinen die Verfahren in dem Sinne, dass der frühere Erfolg von Begutachteten eine gute Vorhersage für den Erfolg in zukünftigen Begutachtungsverfahren ermöglicht. Letzteres wirft die Frage auf, inwiefern im Peer Review ein „old boys network“ am Werk ist und dadurch Entscheidungen nach nicht universellen Kriterien gefällt werden. Gut nachweisbar ist ein solches Defizit in Bezug auf die Nationalität von Begutachteten, die insbesondere bei Zeitschriften als *Bias* deutlich ist. Ob es eine generelle Benachteiligung von jüngeren und weiblichen Forschenden gibt, ist umstritten und deutet daraufhin, dass es eher vor und nach dem *Peer Review*, bspw. durch institutionelle Karrierehindernisse, zu Benachteiligungen kommt.

Aufschlussreicher sind Arbeiten, die eine prozessuale Perspektive einnehmen; die sich also weniger für die Ergebnisse von *Peer Review*-Verfahren interessieren, sondern für deren verfahrensförmigen Ablauf. Baldwin (2018) kann bspw. zeigen, dass die Einführung von externen Gutachtenden in der staatlichen Forschungsförderung eine entscheidende Verfahrensmodifikation war, um politische Unabhängigkeit zu sichern. Lamont zeigt, dass bei interdisziplinär zusammengesetzten *Panels*, ein Aushandlungsprozess stattfindet, in dem fachspezifische Qualitätskriterien durch Fairnesskriterien ergänzt oder gar überlagert werden (Lamont 2009, Mallard et a. 2009). Hirschauer (2005, 2010, 2015) kann zeigen, dass die Funktion von Zeitschriften, die Lesezeit einer Disziplin zu kalibrieren, in eine komplexe kommunikative Interaktion im Peer Review hineinwirkt. Durch ethnografische teilnehmende Beobachtung kann er im Detail nachvollziehen, wie Gutachtende nicht einfach nur ein Urteil abgeben, sondern ihre Einschätzungen in der Kommunikation mit Autor*innen, Herausgeber*innen und anderen Gutachtenden

auf deren Erwartungen ausrichten und in der Interaktion immer wieder (strategisch) anpassen. Als Resultat solcher Arbeiten wird deutlicher, dass Begutachtungsverfahren einerseits nicht einfach Messverfahren für wissenschaftliche Qualität sind und ihre Entscheidungsvalidität problemlos optimiert werden können. Andererseits sind sie intern komplex, weil sie an der Konstruktion und Produktion wissenschaftlicher Qualität mitbeteiligt sind und dadurch Funktionen erfüllen, die über einzelne Verfahrensentscheidungen hinaus reichen.

Bewerten und Autonomie der Wissenschaft

Peer Review ist offensichtlich nicht nur vielfältig, sondern funktioniert auch als Regierungsprinzip der Wissenschaft. Die Allgegenwärtigkeit des Bewertens zeigt sich in der Wissenschaft als eine dezentrale Form des Regierens und Steuerns (Zuckerman & Merton 1971, Jasanoff 1990: 61ff., Weingart 2001: 284ff.). Entscheidungen darüber, welche Forschungsthemen, welche Theorien und Methoden, welche Institutionen und welche Forschenden wichtig und erfolgreich sind, finden verteilt über *Peer Review*-Verfahren an unterschiedlichsten Orten statt (Zeitschriften, Forschungsförderungsorganisationen, Fachgesellschaften, Universitäten, etc.) und entziehen sich damit einer gezielten Steuerung bspw. durch nationalstaatliche Politik. Aus diesem Grund wird in der Literatur oft nicht von Regierung sondern von *Governance* gesprochen. Eine dezentrale Steuerungsform wird in Verbindung damit gebracht, dass der Wissenschaft ein überdurchschnittliches Maß an Autonomie zukommt resp. zukommen sollte (Wilholt 2012). Politisch kommt diese Autonomie in gesetzlichen Sonderregelungen zur Wissenschaftsfreiheit zum Ausdruck (im deutschen Grundgesetz bspw. Paragraph 5: „Forschung und Lehre sind frei.“). Praktisch bedeutet dies, dass die Erwartung besteht, dass Entscheidungen, die die Wissenschaft betreffen, nach wissenschaftlichen Qualitätsmaßstäben getroffen werden sollten, die wiederum in disziplinären Forschungskulturen verankert sein müssen. Da solche Entscheidungen nicht an Akteure außerhalb der Wissenschaft delegiert werden können, da es dort an Fachexpertise mangelt, sind es die beschriebenen Begutachtungsverfahren, in denen nicht nur die Qualität des Wissens gesichert wird, sondern auch die gesellschaftliche Autonomie der Wissenschaft.

Es ist eine historische Besonderheit, dass Bewertungsverfahren und Qualitätsfragen zum zentralen Moment der Steuerung und Regierung von Wissenschaft werden. Die Entstehung der Wissenschaft als gesellschaftlicher Sonderbereich im 16. und 17. Jahrhundert in Europa geht mit der Lösung zweier Probleme einher (Shapin 1994, Biagoli 2002). Einerseits entstehen neue

Formen der Wissensproduktion, die im Empirismus und der Ablehnung traditioneller Wissensautoritäten (Scholastik und Kirche) begründet sind. Diese fordern im Prinzip jeden und jede auf, „im Buch der Natur zu lesen“ und durch unvoreingenommene empirische Betrachtung der Welt neue Erkenntnisse zu produzieren. Dabei ist vorerst unklar, welches neue Wissen als zuverlässig und welche Wissensproduzenten als vertrauenswürdig gelten können. So finden sich in den ersten Ausgaben der *Philosophical Transactions* der *Royal Society* neben Berichten von physikalischen Experimenten ihrer Mitglieder auch aus heutiger Sicht gänzlich unwissenschaftlich oder irrelevant wirkende Berichte bspw. von Handelsreisenden, die über fremde Völker mit Gesichtern zwischen den Schultern berichten. Andererseits wird neues Wissen dringend gebraucht, bspw. im Minenbau, der Schiffsnavigation oder dem Waffenbau, wodurch die Verfügung über dieses Wissen gesellschaftliche Macht verspricht, die aber vorerst nicht unter Kontrolle der herrschenden politischen Autoritäten (König, Kirche) steht. Die gleichzeitige Lösung beider Probleme ist am augenfälligsten in England, wo kraft königlicher Satzung 1660 eine wissenschaftliche Gesellschaft (*Royal Society*) gegründet wird, deren Mitglieder dem Adelsstand angehören.[2] Als „peers“ (wörtlich: gleich, ebenbürtig; meint aber: hochadlig) stellen diese in staatstragender Funktion sicher, dass das publizierte Wissen den politischen Status Quo nicht gefährdet und werden dafür von der Zensur ausgenommen, um wissenschaftlich publizieren zu können. Die Kontrolle des zu Publizierenden geschieht dann im Kreis der *Peers*, die Zuschriften begutachten und dabei auch gleich auf ihre wissenschaftliche Qualität prüfen; deshalb *Peer Review*.

Auch wenn sich ganz sicher keine gradlinige historische Entwicklung aus dem 17. Jahrhundert in die Gegenwart verfolgen lässt, so kann man doch sagen, dass diese doppelte Leistung, auf der die relative Autonomie der Wissenschaft gründet – Qualitätssicherung nach Innen und Legitimierung nach Außen – weiterhin erbracht wird. Dies geschieht nicht mehr durch eine einzelne Fachgesellschaft wie die *Royal Society* mit privilegierten Beziehungen zu politischen Autoritäten, sondern durch eine ganze Landschaft von Bewertungspraktiken. Zwei Entwicklungen können verdeutlichen, wie sich diese relative gesellschaftliche Autonomie gegenwärtig darstellt:

Erstens ist die Verbreitung und Formalisierung von *Peer Review*-Verfahren eine Reaktion auf Entwicklungen in den 1960er und 1970er Jahren, um der Politisierung der Forschungs-

2 Die Gründungsgeschichte der Royal Society hat in der Wissenschaftsforschung viel Aufmerksamkeit erhalten, weil sie als exemplarischer Fall für die Institutionalisierung moderner Wissenschaft im 17. Jahrhundert gesehen werden kann. Zu Fragen des Bewertens von Forschung liefern die Arbeiten von Zuckerman & Merton (1971) und Shapin (1994) einen guten Einstieg.

förderung durch parteipolitische Interessen entgegen zu wirken (Baldwin 2018). David Guston (2000) spricht davon, dass eine Art Gesellschaftsvertrag entstanden ist, der die öffentliche Finanzierung von Wissenschaft an allgemeine Rechenschaftspflichten gegenüber der Politik bindet, aber ohne dass dadurch Einfluss auf Themen und Inhalte der Forschung stattfinden sollen. Trotzdem ist seit den 1980er Jahren eine zunehmend an Themen und Inhalten interessierte Wissenschaftspolitik zu beobachten, die von der Forschung bspw. mehr gesellschaftliche Relevanz oder mehr Interdisziplinarität einfordert (Flink & Kaldewey 2018). Als Folge tauchen diese Forderungen einerseits zunehmend als thematische Setzungen für Programme der Forschungsförderung auf. Programme, die gezielt nur interdisziplinäre, translationale (sprich: anwendungsbezogene) oder gar Forschung mit nichtwissenschaftlichen Akteuren (*citizen science*) fördern, sind ein Instrument, mit dem öffentlichen und politischen Rechenschaftsforderungen nachgekommen wird. Andererseits tauchen diese Themen vermehrt auch als Qualitätskriterien im *Peer Review* selbst auf, wo sie mehr oder weniger gleichberechtigt neben fachnahen Kriterien wie methodischer Strenge oder Originalität bei der Bewertung zur Anwendung kommen.

Zweitens ist der gesamtgesellschaftliche Trend zur vermehrten Evaluation („audit society“, Power 1997) auch in der Wissenschaft zu beobachten, wo er in Kombination mit digitalen Formen der Kommunikation und sozialer Medien neue Bewertungsphänomene produziert. Ein Trend zur vermehrten Evaluation in der Wissenschaft heißt, dass neben den allgegenwärtigen und informellen Praktiken des Bewertens eine Zunahme an formalisierten Evaluationsverfahren zu verzeichnen ist. Hochschulrankings sind hierfür ein gutes Beispiel, weil sie einerseits ein außerwissenschaftliches Publikum adressieren und andererseits nur peripher auf fachbezogene Expertise für die Bewertung angewiesen sind (Hazelkorn 2014). Grundlage hierfür sind quantifizierbare Informationen über die Wissenschaft, die dann als statistisch nutzbare Indikatoren verwendet werden können. Für Hochschulrankings sind das bspw. die Anzahl hochzitierter Publikationen einer Universität. Viele dieser quantifizierbaren Informationen sind Metadaten wissenschaftlicher Kommunikation, die wertend gedeutet werden können; so wenn Zitate als Einfluss, Reputation oder Qualität gedeutet werden. Da wissenschaftliche Kommunikation in der Zwischenzeit mehrheitlich digital stattfindet, entstehen auch hier in großem Umfang Metadaten die sich für Bewertungen nutzen lassen (Gauch, in diesem Band). *Tweets*, *Downloads*, *Views*, etc. sind Bestandteil sog. alternativer Metriken (*altmetrics*), die für die Bewertung von Forschungsarbeiten, Forschenden und Forschungsorganisationen herangezogen werden und die die Bedutung von digitalen Kom-

munikationsplattformen (*social media*) im wissenschaftlichen Alltag zum Ausdruck bringen (Sugimoto et al. 2017).

Bewerten zur Steuerung und Regierung der Wissenschaft

Wie oben beschrieben zieht sich der organisierte Skeptizismus durch unzählige informelle Praktiken und formalisierte Verfahren des Bewertens, die nach Innen auf unterschiedlichste Qualitätsanforderungen einer nach Disziplinen differenzierten Wissenschaft Rücksicht nehmen und die nach Außen für unterschiedlichste Anspruchsgruppen in multipolaren Wissensgesellschaften Anschlüsse schaffen. Dass die Frage „Is it peer reviewed?“ zunehmend auch im massenmedialen Diskurs auftaucht, belegt die Schlüsselstelle, die Bewertungsverfahren dabei einnehmen, täuscht aber auch über die Vielfältigkeit dieser Landschaft hinweg. Es ist eben nicht nur das *Peer Review*, sondern eine ganze Reihe von Institutionen, die dies leisten. Zwei davon sollen hier exemplarisch herausgegriffen werden: Grenzorganisationen und soziale Bewegungen.

Guston (2001) bezeichnet jene Organisationen als Grenzorganisationen, die zwischen der Wissenschaft und anderen gesellschaftlichen Akteuren vermitteln. Staatliche Forschungsförderungsorganisationen sind hierfür ein klassisches Beispiel, weil sie gegenüber der Wissenschaft für die Bereitstellung und Verteilung öffentlicher Forschungsmittel sorgen und gegenüber der Politik für Rechenschaft über deren Verwendung. Dass dies mehrheitlich über *Peer Review* als Bewertungsverfahren geschieht, ist dabei nicht zwingend, wie man an anderen Grenzorganisationen sehen kann. Guston erwähnt bspw. Organisationen, die geschaffen wurden, um Fällen wissenschaftlichen Fehlverhaltens nachzugehen, in den USA bspw. das *Office of Research Integrity*. Deren Bewertungsverfahren sind eher dem Rechtssystem entliehen und resultieren in Sanktionen gegen inkriminierte Forschende. Ihre Grenzfunktion besteht darin, ein Minimum an Forschungsqualität nach Innen abzusichern und nach Außen für Legitimität zu sorgen.

Während Grenzorganisationen, wie der Name sagt, Grenzarbeit leisten, gibt es in dieser Bewertungslandschaft auch zunehmend soziale Bewegungen, die über die Grenze der Wissenschaft hinweg aktiv sind und Fragen der Forschungsqualität adressieren. Als soziale Bewegungen werden sie hier bezeichnet, weil sie von einer Unzufriedenheit bzgl. der Qualität von Forschung ausgehen, die dann zu kollektiven Aktivitäten von einzelnen Forschenden führt, die vorerst nur temporär und minimal organisiert sein können (Hess et al. 2008). Austausch und Koordination findet dann vermehrt über soziale Medien und

durch den Einsatz algorithmischer Tools statt. Ein in Deutschland prominentes Beispiel ist *VroniPlag Wiki* (https://vroniplag.fandom.com/), wo Plagiate in Doktorarbeiten aufgedeckt und dokumentiert werden (Hesselmann & Reinhart 2020). Die Zusammenarbeit der mehrheitlich anonymen Gruppe findet offen über eine Wiki-Plattform statt und die Ergebnisse der oft minutiösen Dokumentation von Plagiaten, erregen immer wieder großes mediales Interesse; nicht zuletzt weil so auch Personen des öffentlichen Lebens, insb. aus der Politik, überführt wurden. Dieses Bewertungsverfahren (Aufdecken von Plagiaten) problematisiert nicht nur ein anderes bestehendes Verfahren in der Wissenschaft (Promotion), sondern involviert auch nichtwissenschaftliche Akteure in der Bewertung von Forschungsqualität. Weitere Beispiele für soziale Bewegungen zu Forschungsqualität finden sich im letzten Abschnitt.

Zusammengefasst ergibt sich folgendes Bild: Das Bewerten von wissenschaftlicher Leistung und Qualität ist allgegenwärtig und vielfältig, so dass man bildlich sagen könnte, dass das Feld der Wissenschaft durch eine Landschaft von Bewertungspraktiken strukturiert ist. Diese Bewertungspraktiken sind auf zwei primäre Ziele ausgerichtet, einerseits auf die Sicherung einer minimalen Qualität und andererseits auf die Auszeichnung besonders wertvoller Forschung. Obwohl viele Bewertungspraktiken alltäglich und informell ablaufen, sind es vor allem formalisierte Bewertungs*verfahren*[3], wie das *Peer Review*, denen eine Steuerungs- oder Regierungsfunktion zukommen. Diese Steuerungsfunktion regelt innerhalb der Wissenschaft, welche Forschung finanziert und publiziert wird, welchen Arbeiten und Akteuren Aufmerksamkeit und Reputation zukommt. Gleichzeitig regeln diese Verfahren aber auch das Außenverhältnis der Wissenschaft zur Gesellschaft, indem sie eine Grenze um das Wissenschaftliche ziehen und den Wert wissenschaftlicher Arbeit nach Außen legitimieren. Diese Innen- und Außenverhältnisse sind dynamisch, was sich abschließend an aktuellen Entwicklungen darstellen lässt.

Aktuelle Entwicklungen

Eine derart vielfältige Landschaft der Bewertungsverfahren zeugt einerseits davon, dass die relative Autonomie der Wissenschaft weiterhin über Grenzarbeit und Grenzorganisationen abgesichert wird. Andererseits zeugen soziale Bewegungen, bspw. zur Plagiatsaufdeckung oder zur Replikation von Forschungsergebnissen davon, dass diese Abgrenzung in gegenwärtigen Wissensgesellschaften nur bedingt möglich ist. Es eröffnen sich daraus gegenwärtig zwei paradoxale Herausforderungen

3 Zur Verfahrensförmigkeit des Bewertens in der Wissenschaft siehe Schendzielorz & Reinhart 2020: 103-106.

für die Bewertung wissenschaftlicher Qualität: 1. Je offener und zugänglicher die Wissenschaft wird, desto mehr erlaubt diese Offenheit (fachfremde) Kritik an der Qualität der Forschung. 2. Wissenschaft ist mit konfligierenden Ansprüchen an Wertfreiheit konfrontiert: Sie soll Wissen und Expertise produzieren, die ihre gesellschaftliche Autorität dadurch begründet, dass sie unabhängig von (politischen) Interessenlagen ist. Sie soll aber auch gesellschaftlich relevante Werte (Nachhaltigkeit, Diversität, etc.) in ihre Bewertungsverfahren aufnehmen, während immer nachdrücklicher kritisiert wird, dass diese Bewertungsverfahren den schon bestehenden Ansprüchen an Universalität und Offenheit nicht gerecht würden. Es bleibt vorerst offen, wie diese Herausforderungen tatsächlich bewältigt werden, aber es lassen sich zumindest Beispiele anführen, wo diese gegenwärtig problematisiert und bearbeitet werden.

Ausgehend von der biomedizinischen Forschung tauchen zunehmend Fragen nach Mindeststandards von Forschungsqualität auf. „Why most published research findings are false“ (Ioannidis 2005) lautet der unironische Titel eines vielbeachteten Artikels, an den sich Fragen anschließen, ob die publizierten Arbeiten denn überhaupt alle replizierbar seien, ob die statistischen Analysen genügend zuverlässig seien (p-hacking) oder ob nur selektiv jene Experimente publiziert würden, die erfolgreich ein Phänomen bestätigen können (file-drawer problem). Als Reaktion darauf gibt es über die Medizin hinaus sog. Replikationsinitiativen (Stroebe & Strack 2014, Reinhart 2016), die die für die Reputation undankbare Aufgabe übernehmen, wichtige Forschungsarbeiten zu wiederholen, um sie mit größerer Sicherheit bestätigen zu können. Zeitschriften haben begonnen, die Autor*innen aufzufordern, nicht nur die finalen Texte mit den Forschungsergebnissen zu publizieren, sondern auch die zugrundeliegenden Daten und die Protokolle zu deren Analyse öffentlich zu hinterlegen. Schließlich werden auch digitale Tools entwickelt, die große Mengen wissenschaftlicher Publikationen auf Qualität prüfen können (Introna 2015, Weber-Wulff 2019): Plagiats- und Bildmanipulationserkennung, Erkennung von gängigen statistischen Fehlern oder Überprüfung der Verfügbarkeit zugehöriger Daten. Viele dieser Initiativen bilden den Kern von kleineren oder größeren sozialen Bewegungen innerhalb der Wissenschaft und zielen darauf, dass die entsprechenden Tools und Kriterien auch in den Bewertungsverfahren bei Zeitschriften oder in der Forschungsförderung Eingang finden.

Begründet werden diese Qualitätssicherungsmaßnahmen oft damit, dass mehr Anreize geschaffen werden sollen, hochwertige Forschung zu produzieren, weil dies direkt in Bewertungsverfahren belohnt würde. Diese naive Sichtweise trifft auf eine gleichzeitig stattfindende Diskussion, die die systemischen Effekte bestehender Evaluationssysteme problematisiert. Dort

geht es um die Frage, ob die Wissenschaft nicht mit formalisierten Bewertungsverfahren überladen sei und sowohl die Begutachteten wie die Begutachtenden überfordere (Neidhardt 2016: 269f.). Gutachtende seien überlastet und es sei deshalb schwierig, die notwendige Fachexpertise in den Verfahren aufzubringen. Die Forschenden würden durch unablässigen Evaluationsdruck und unsichere Karriereperspektiven demotiviert (Loveday 2018). Wobei hinzukommt, dass dieser Druck durch soziale Medien wie *ResearchGate*, *Google Scholar*, etc. noch verschärft würde (Reinhart 2021). Zerschlagen ist in der Zwischenzeit auch die Hoffnung, dass eine Entlastung dadurch geschaffen werden kann, dass ein Teil des als aufwändig wahrgenommenen *Peer Review* durch die Nutzung von quantitativen Indikatoren ersetzt werden kann. Die Ausweitung der Praxis, Forschende, Zeitschriften oder Universitäten nach berechenbaren Indikatoren (h-index, impact factor, altmetrics, Drittmitteleinwerbungen, Studierendenzahlen, etc.) zu bewerten (Blümel & Gauch 2021), hat im Gegenzug zu Aufrufen, Manifesten und gar Boykotten gegen eine solche Praxis geführt (DORA Declaration: https://sfdora.org/, Leckert 2021). Es zeichnet sich ab, dass die Bewertung wissenschaftlicher Qualität zu einem zentralen wissenschaftspolitischen Thema wird. Forschungsorganisationen, nationale Wissenschaftspolitiken und internationale Forschungsförderung sind aufgerufen, Evaluationssysteme zu gestalten, die auf „intrinsic merits“ und einen „responsible use of quantitative indicators“ ausgerichtet sind (Europäische Kommission 2021: 3). Es wird sich zeigen müssen, inwiefern dieser Gestaltungswille den komplexen Anforderungen gerecht werden kann: Die Regierung der Wissenschaft bleibt eng mit der Bewertung von Forschungsqualität verknüpft, aber die Vervielfältigung von gesellschaftlichen Ansprüchen strapaziert sowohl die bestehenden als auch neue Verfahren, so dass die relative gesellschaftliche Autonomie der Wissenschaft ein Dauerthema bleiben wird.

Lektüreempfehlungen

(1) Melinda Baldwin (2018) identifiziert im Aufsatz "Scientific Autonomy, Public Accountability, and the Rise of Peer Review in the Cold War United States" den historischen Moment, in dem der Einsatz von Peer Review selbstverständlich wurde. Kritik an der staatlichen Forschungsförderung wird seit den 1970er Jahren mit der Einführung oder Verbesserung von Peer Review begegnet.

(2) Mario Biagioli (2002) zeigt in "From book censorship to academic peer review" wie das Peer Review in seinen historischen Anfängen im 17. Jahrhundert gleichermaßen als Instrument der Qualitätssicherung und der Zensur entsteht.

(3) Stefan Hirschauer (2005) schöpft aus teilnehmender Beobachtung bei einer soziologischen Zeitschrift und beschreibt in "Publizierte Fachurteile. Lektüre und Bewertungspraxis im Peer Review" detailliert was während des Begutachtungsprozesses vor sich geht. Es werden nicht einfach Urteile über Manuskripte abgeben, sondern es entfalten sich komplexe Muster strategischer Kommunikation zwischen Autor*innen, Gutachter*innen und Herausgeber*innen, bei denen sich die Akteure auch gegenseitig beobachten und bewerten.

(4) Vik Loveday (2018) hat für den Aufsatz "Luck, chance, and happenstance? Perceptions of success and failure amongst fixed-term academic staff in UK higher education" Forschende danach gefragt, wie sie sich ihre Erfolge und Misserfolge erklären. Aus einem geschickten Befragungsdesign kann sie ableiten, dass die Unsicherheit befristeter Stellen und regelmäßige Evaluationen zu paradoxalen Selbsteinschätzungen führen: Eigene Erfolge werden als glücklich und eigene Misserfolge als verdient interpretiert, während die Erfolge anderer als verdient und die Misserfolge als unglücklich gedeutet werden.

Baldwin, M. (2018) 'Scientific Autonomy, Public Accountability, and the Rise of "Peer Review" in the Cold War United States', Isis, 109(3), pp. 538–558.

Biagioli, M. (2002) 'From book censorship to academic peer review', Emergences: Journal for the Study of Media & Composite Cultures, 12(1), pp. 11–45.

Blümel, C. und Gauch, S. (2020) 'History, Development and Conceptual Predecessors of Altmetrics', in Handbook Bibliometrics. De Gruyter Saur, pp. 191–200.

Boltanski, L. and Thévenot, L. (1999) 'The Sociology of Critical Capacity', European Journal of Social Theory, 2(3), pp. 359–377.

Braun, D. (1997) Die politische Steuerung der Wissenschaft.: Ein Beitrag zum kooperativen Staat. Frankfurt am Main: Campus.

Dahler-Larsen, P. (2012) The evaluation society. Stanford, California: Stanford Business Books.

Dahler-Larsen, P. (2019) Quality: from Plato to Performance. Springer.

Directorate-General for Research and Innovation (European Commission) (2021) Towards a reform of the research assessment system: scoping report. LU: Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2777/707440 (Zugriff: 11. März 2022).

Flink, T. und Kaldewey, D. (2018) 'The Language of Science Policy in the Twenty-First Century', in Kaldewey, D. und Schauz, D. (Hrsg.) Basic and Applied Research: The Language of Science Policy in the Twentieth Century. Berghahn Books (European Conceptual History), pp. 251–284.

Guston, D.H. (2000) Between Politics and Science: Assuring the Integrity and Productivity of Research. Cambridge University Press.

Guston, D.H. (2001) 'Boundary Organizations in Environmental Policy and Science: An Introduction', Science, Technology, & Human Values, 26(4), pp. 399–408.

Guthrie, S., Ghiga, I. and Wooding, S. (2018) 'What do we know about grant peer review in the health sciences?', F1000Research, 6: 1335.

Hazelkorn, E. (2014) 'Reflections on a Decade of Global Rankings: what we've learned and outstanding issues', European Journal of Education, 49(1), pp. 12–28.

Hess, D. et al. (2008) ‘Science, Technology, and Social Movements’, in Hackett, E.J. et al. (Hrsg.) The handbook of science and technology studies. The MIT Press, pp. 473–498.
Hesselmann, F. und Reinhart, M. (2020) ‘Fragmentierte Sichtbarkeiten: Visualität, Sichtbarkeit und Unsichtbarkeit beim Umgang mit wissenschaftlichem Fehlverhalten’, Kriminologisches Journal, 52(1), pp. 6–20.
Hirschauer, S. (2004) ‘Peer Review Verfahren auf dem Prüfstand: Zum Soziologiedefizit der Wissenschaftsevaluation’, Zeitschrift für Soziologie, 33(1), pp. 62–83.
Hirschauer, S. (2005) ‘Publizierte Fachurteile. Lektüre und Bewertungspraxis im Peer Review’, Soziale Systeme, 11(1), pp. 52–82.
Hirschauer, S. (2010) ‘Editorial Judgments: A Praxeology of “Voting” in Peer Review’, Social Studies of Science, 40(1), pp. 71–103.
Hirschauer, S. (2015) ‘How Editors Decide. Oral Communication in Journal Peer Review’, Human Studies, 38(1), pp. 37–55.
Hirschman, A.O. (1970) Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States. Harvard University Press.
Hug, S.E. und Aeschbach, M. (2020) ‘Criteria for assessing grant applications: a systematic review’, Palgrave Communications, 6(1), p. 37. doi:10/ggnqm8.
Introna, L.D. (2016) ‘Algorithms, Governance, and Governmentality: On Governing Academic Writing’, Science, Technology, & Human Values, 41(1), pp. 17–49.
Ioannidis, J.P.A. (2005) ‘Why Most Published Research Findings Are False’, PLOS Med, 2(8), p. e124.
Jasanoff, S. (1990) The Fifth Branch. Science Advisers as Policymakers. Harvard University Press.
Krüger, A.K. und Reinhart, M. (2016) ‘Wert, Werte und (Be)Wertungen. Eine erste begriffs- und prozesstheoretische Sondierung der aktuellen Soziologie der Bewertung’, Berliner Journal für Soziologie, 26(3–4), pp. 485–500.
Krüger, A.K. und Reinhart, M. (2017) ‘Theories of Valuation - Building Blocks for Conceptualizing Valuation Between Practice and Structure’, Historical Social Research, 42(1), pp. 263–285.
Lamont, M. (2012) ‘Towards a Sociology of Valuation: Convergence, Divergence, and Synthesis’, Annual Review of Sociology, 38(1), pp. 201–221.
Leckert, M. (2021) ‘(E-) Valuative Metrics as a Contested Field: A Comparative Analysis of the Altmetrics- and the Leiden Manifesto’, Scientometrics, 126(12), pp. 9869–9903.
Loveday, V. (2018) ‘“Luck, chance, and happenstance? Perceptions of success and failure amongst fixed-term academic

staff in UK higher education"', The British Journal of Sociology, 69(3), pp. 758–775.
Mallard, G., Lamont, M. und Guetzkow, J. (2009) 'Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review', Science Technology & Human Values, 34(5), pp. 573–606.
Merton, R.K. (1973) 'The Normative Structure of Science', in The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, pp. 267–279.
Mirowski, P. (2018) 'The future(s) of open science', Social Studies of Science, 48(2), pp. 171–203.
Möller, Torger, Antony, Philipp, Hinze, Sybille & Hornbostel, Stefan (2012) Exzellenz begutachtet. Befragung der Gutachter in der Exzellenzinitiative. Berlin (iFQ-Working Paper, 11).
Neidhardt, F. (2016) 'Selbststeuerung der Wissenschaft durch Peer-Review-Verfahren', in Simon, D. et al. (Hrsg.) Handbuch Wissenschaftspolitik. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 261–277.
Potthast, J. (2017) 'The Sociology of Conventions and Testing', in Benzecry, C.E., Krause, M., und Ariail Reed, I. (Hrsg.) Social Theory Now. Chicago University Press, pp. 337–360.
Power, M. (1997) The Audit Society: Rituals of Verification. Oxford University Press.
Reinhart, M. (2012) Soziologie und Epistemologie des Peer Review. Baden-Baden: Nomos.
Reinhart, M. (2016) 'Rätsel und Paranoia als Methode - Vorschläge zu einer Innovationsforschung der Sozialwissenschaften', in Froese, A., Simon, D., und Böttcher, J. (Hrsg.) Sozialwissenschaften und Gesellschaft: Neue Verortungen von Wissenstransfer. Bielefeld: Transcript, pp. 159–191.
Reinhart, M. (2021) 'Open Science as an Engine of Anxiety: How Scientists Promote and Defend the Visibility of Their Digital Selves, while Becoming Fatalistic about Academic Careers'. SocArXiv. doi:10.31235/osf.io/2vr7j.
Reinhart, M. und Schendzielorz, C. (2021a) 'Trends in Peer Review'. SocArXiv. doi:10.31235/osf.io/nzsp5.
Reinhart, M. und Schendzielorz, C. (2021b) 'Peer Review Procedures as Practice, Decision, and Governance – Preliminaries to Theories of Peer Review'. SocArXiv. doi:10.31235/osf.io/ybp25.
Schendzielorz, C. und Reinhart, M. (2020) 'Die Regierung der Wissenschaft im Peer Review / Governing Science Through Peer Review', dms – der moderne staat – Zeitschrift für Public Policy, Recht und Management, 13(1), pp. 101–123.
Shapin, S. (1994) A Social History of Truth: Civility and Science in Seventeenth-Century England. University Of Chicago Press.
Simmel, G. (2008) Philosophie des Geldes. Frankfurt am Main: Suhrkamp.

Stroebe, W. und Strack, F. (2014) ‘The Alleged Crisis and the Illusion of Exact Replication’, Perspectives on Psychological Science, 9(1), pp. 59–71.
Sugimoto, C.R. et al. (2017) ‘Scholarly use of social media and altmetrics: A review of the literature’, Journal of the Association for Information Science and Technology, 68(9), pp. 2037–2062.
Weber-Wulff, D. (2019) ‘Plagiarism detectors are a crutch, and a problem’, Nature, 567(7749), p. 435.
Weingart, P. (2001) Die Stunde der Wahrheit?: Zum Verhältnis der Wissenschaft zu Politik, Wirtschaft und Medien in der Wissensgesellschaft. Weilerswist: Velbrück Wissenschaft.
Wilholt, T. (2010) Die Freiheit der Forschung, Suhrkamp.
Zuckerman, H. and Merton, R.K. (1971) ‘Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System’, Minerva, 9(1), pp. 66–100.