Michael Ochsner

16 Messung von Forschungsleistungen? Was gemessen wird und was gemessen werden will

16.1 Einleitung

Öffentliche Institutionen sind seit den letzten Jahrzehnten erhöhtem Effizienzdruck und der Forderung nach direkter Rechenschaftsablegung ausgesetzt. Diese Entwicklung machte auch nicht vor den Toren der Universität halt (Hamann, 2016; Hammarfelt und De Rijcke, 2014; Kekäle, 2002; Mali, Pustovrh, Platinovsek, Kronegger und Ferligoj, 2017; Schmidt, 2010), was ihren Niederschlag in unzähligen Reformen nicht zuletzt bezüglich der Forschungsevaluation fand. So entwickelten Universitäten komplexe Evaluationssysteme für Forschungsleistungen (siehe z. B. Geuna und Martin, 2003; Ochsner, Kulczycki und Gedutis, 2018).[65] Solche Systeme basierten aufgrund der leichten Verfügbarkeit von quantitativen Indikatoren – angeboten von Diensten wie Web of Science (ehemals ISI, dann Thomson Reuters, heute Clarivate Analytics) oder Scopus (Elsevier) – und der Vorliebe von Managern für Vergleiche einfacher Zahlen, die als objektiver und vergleichbarer als Experteneinschätzungen oder komplexe Indikatorensysteme gesehen werden, primär auf sogenannten Forschungsindikatoren, wie zum Beispiel der Anzahl an Publikationen oder der Anzahl an Zitationen. Zunächst waren die Natur- und Lebenswissenschaften wegen der hohen Kosten der Infrastruktur, die häufig für natur- und lebenswissenschaftliche Projekte vonnöten ist, von solchen Evaluationssystemen betroffen. In den letzten Jahren rückten auch die Sozial- und Geisteswissenschaften (SGW) in den Fokus solcher Evaluationssysteme (Burrows, 2012; Guillory, 2005). Bald jedoch zeigte sich, dass gerade bezüglich der SGW große Probleme der Interpretierbarkeit der gängigen quantitativen Forschungsindika-

65 Evaluationen an Hochschulen finden häufig für Forschung und Lehre getrennt statt, respektive wird unter „institutioneller Evaluation" häufig die Evaluation von Lehre ausgespart oder stiefmütterlich behandelt. Dieser Beitrag bespricht Kriterien für Evaluation an Hochschulen, was Lehre nicht ausschließt, jedoch primär Forschungsleistungen berücksichtigt. Dies heißt allerdings nicht zwingend, dass in der Evaluation von Forschungsleistungen die Lehre nicht einbezogen wird, insbesondere in den Sozial- und Geisteswissenschaften wird die Lehre bisweilen als ein relevanter Punkt angeführt (siehe etwa Bozkurt Umur, Diaz-Bone und Surdez, 2017; Hug, Ochsner und Daniel, 2013; Ochsner, Hug und Daniel, 2012a).

Michael Ochsner, FORS – Swiss Centre of Expertise in the Social Sciences, Lausanne
ETH Zürich

https://doi.org/10.1515/9783110689884-017

toren bestehen: So sind die Datenbanken, auf deren Grundlage die Forschungsindikatoren erhoben werden, nicht auf die SGW abgestimmt. Die relevante Literatur wird nicht erfasst, und Publikations- und Zitationsdaten sind somit weder komplett noch valide (vgl. z. B. van Leeuwen, 2013). Dies hat verschiedene Gründe: Zu den wichtigsten gehören die Sprache, die Publikationsformen und epistemologische Unterschiede zu den Natur- und Lebenswissenschaften (vgl. z. B. Hicks, 2004; Nederhof, 2006; Lack, 2008; van Leeuwen, 2013). Während die gängigen Datenbanken sich auf englischsprachige Literatur beschränken, ist in vielen geistes- und sozialwissenschaftlichen Fächern die lingua franca nicht Englisch und bisweilen ist die englischsprachige Literatur sogar marginal in ihrer Bedeutung. Die Datenbanken basieren primär auf Zeitschriftenartikel, während die Publikationsformen in den SGW deutlich vielfältiger sind. Die wichtigsten Forschungsbefunde werden in vielen Fächern in Büchern publiziert; Museumskataloge, Rapporte und andere Publikationsformen, die die Gesellschaft anvisieren, sind weitere relevante Publikationsformen, die nicht in den Datenbanken aufgenommen werden. Schließlich – und fundamentaler – funktioniert aber die Wissensgenerierung anders als in den Natur- und Lebenswissenschaften. Zitationen werden anders eingesetzt, und die Wissensgenerierung folgt keinem linearen Fortschrittsgedanken. Vielmehr geht es um die Erweiterung der Interpretationen und des Erkenntnisraums, weshalb gleichzeitig verschiedene Paradigmen koexistieren. Somit ist auch der Zeithorizont, innerhalb welchem Forschung relevant ist, ein anderer. Alle diese Gründe führen dazu, dass die gängigen Forschungsindikatoren in den SGW Forschungsleistungen nicht adäquat abbilden können.

Aber auch in den Natur- und Lebenswissenschaften wird in letzter Zeit die Kritik an Evaluation mittels rein quantitativer Verfahren, insbesondere der Zitationsanalysen, immer lauter (Lawrence, 2002; MacRoberts und MacRoberts, 2017; Molinié und Bodenhausen, 2010). Offensichtlich besteht ein Mangel an adäquaten Instrumenten zur Beurteilung von Forschungsleistungen. Während sich die Natur- und Lebenswissenschaften ihrer Natur gemäß auf die *Messung* von Forschungsleistungen konzentriert haben, wählten die SGW ihrer Natur entsprechend in den letzten Jahren den Weg der *Kritik und Reflexion* der Forschungsevaluation und ihrer Methoden. So entstanden viele Projekte, die nach alternativen Methoden suchten (für einen Überblick, siehe Ochsner, Hug und Galleron, 2017). Im Zentrum solcher Projekte stand oft auch die Frage danach, was denn überhaupt gemessen werden soll. Dieser Beitrag befasst sich mit ebendiesem Thema: Was soll in den Forschungsevaluationen denn gemessen werden? Welche Kriterien für Forschungsqualität lassen sich finden? Welche dieser Kriterien sind quantitativ messbar, welche nicht?

Dieser Beitrag ist folgendermaßen gegliedert: Zuerst wird ein kurzer Überblick verschiedener nationaler Evaluationssysteme in Europa gegeben. Anschließend wird auf die Problematik des Messgegenstands eingegangen und erläutert, wie idealerweise gemessen werden sollte. Dann wird anhand von Kriterien in nationalen Evaluationsprozeduren und empirisch hergeleiteten Kriterienkatalogen aufgezeigt, wie kom-

plex und multidimensional Forschungsqualität ist. Schließlich wird anhand der Geisteswissenschaften gezeigt, wie Forschungsqualität operationalisiert werden kann.

16.2 Nationale Evaluationssysteme

Die Evaluationspraktiken unterscheiden sich stark zwischen Ländern und gar Universitäten. In den letzten Jahren sind deshalb verschiedene Typologien vorgeschlagen worden (Coryn et al., 2007; Hicks, 2010, 2012; Geuna und Martin, 2001, 2003; Jonkers und Zacharewicz, 2016; Lepori, Reale und Spinello, 2018; von Tunzelmann und Kraemer Mbula, 2003; Zacharewicz, Lepori, Reale und Jonkers, 2018). Oft werden dabei zwei grundsätzliche Typen von Forschungsevaluation unterschieden: leistungsabhängige Systeme (manchmal auch summative Systeme genannt), also solche, die finanzielle Konsequenzen nach sich ziehen, und formative Systeme, die eine reflexive Weiterentwicklung zum Ziel haben, d. h. Systeme, in denen es primär um eine Standortbestimmung und um die Erörterung von Möglichkeiten zur Verbesserung und Entwicklung der evaluierten Einheit geht. So einfach lassen sich nun aber Länder nicht in diese zwei Kategorien einteilen, insbesondere weil sich die Prozesse innerhalb der Gruppen so stark unterscheiden, dass sich auch Experten in Forschungsevaluation nicht darüber einig sind, wie das System im eigenen Land eingestuft werden soll (Galleron, Ochsner, Spaapen und Williams, 2017; Ochsner, Kulczycki und Gedutis, 2018). Bisweilen sind in einem Land sowohl leistungsabhängige wie auch formative Komponenten implementiert (z. B. Norwegen). Die meisten Typologien bleiben zudem deskriptiv und nehmen nur eine Evaluationsprozedur pro Land in den Blick, während sich in allen Ländern komplexe Evaluationssysteme etabliert haben, die aus verschiedenen Prozeduren bestehen. Eine aktuelle Typologie, die die Evaluationssysteme anhand einer breiten Auswahl von Merkmalen für viele Länder beschreibt (Galleron et al., 2017; Ochsner et al., 2018), zeigt, dass fast jedes Land sein individuelles Evaluationssystem hat (siehe auch für die Rechtswissenschaften van Gestel und Lienhard, 2019). Nichtsdestotrotz lassen sich grob zwei Dimensionen feststellen, anhand derer fünf Idealtypen von Evaluationssystemen ausgemacht werden können (siehe Abbildung 16.1).

Die erste Dimension teilt die europäischen Länder danach auf, ob sie eine nationale bibliografische Datenbank haben, ob die Evaluationen primär national organisiert sind und ob Metriken eine größere Rolle spielen. Die zweite Dimension umfasst die Anpassung des Evaluationssystems an fachspezifische Eigenheiten, insbesondere, ob für die Geisteswissenschaften angemessene Evaluationsmethoden verwendet werden. Die fünf Idealtypen lassen sich folgendermaßen umschreiben: 1) „keine nationale Datenbank, keine metrische Evaluation, keine Anpassung für SGW (oder nicht fachspezifisch)“ mit Zypern (CY), Frankreich (FR), Island (IS), Mazedonien (MK), Malta (MT), Montenegro (ME), Portugal (PT) und Spanien (ES); 2) „qualitative Beurteilung, Anpassung für SGW (oder fach-spezifisch)“ mit Österreich (AT), Deutschland (DE),

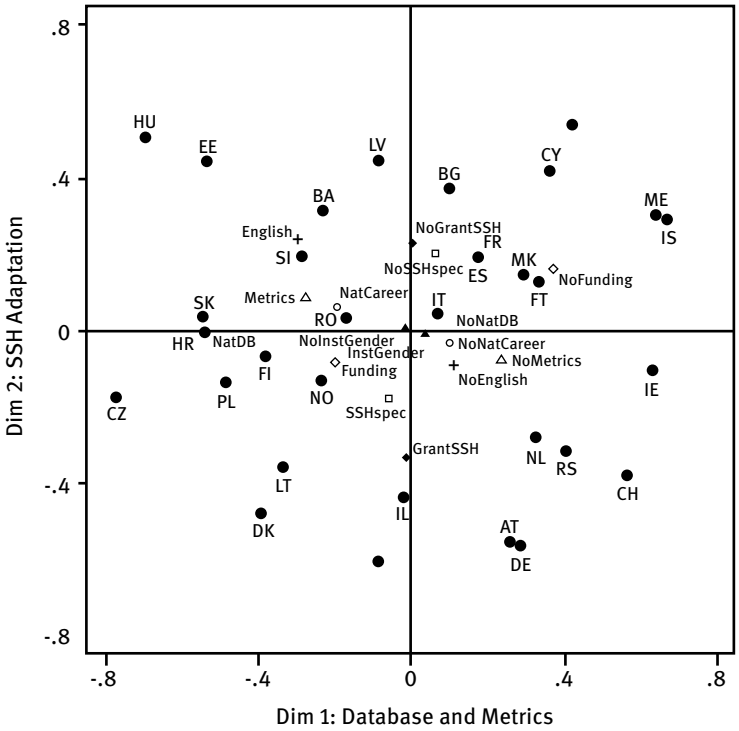

Abb. 16.1: Nationale Evaluationssysteme für Forschung (Quelle: Ochsner et al., 2018).
Bemerkungen: Kreise repräsentieren Länder, alle anderen Symbole stehen für die Dummy-Variablen der Charakteristika von den Evaluationssystemen. English/NoEnglish: das System setzt Anreize auf Englisch zu publizieren; (No)Funding: die Resultate der Evaluation haben finanzielle Konsequenzen; (No)GrantSSH: GSW-spezifische Projektförderung vorhanden; (No)InstGender: Evaluationsprozeduren berücksichtigen geschlechtsspezifische Aspekte; (No)Metrics: Hauptmethode sind Metriken; (No)NatCareer: Nationale Institution für Promotionen und Habilitationen; (No)NatDB: nationale Publikationsdatenbank; (No)SSHspec: GSW-spezifische Evaluationsprozedur.

Irland (IE), den Niederlanden (NL), Serbien (RS) und der Schweiz (CH); 3) „qualitative Beurteilung, mit leistungsabhängiger Komponente“ mit Litauen (LT), Norwegen (NO), Südafrika (ZA); 4) „quantitative Beurteilung, leistungsabhängig“ mit der Tschechischen Republik (CZ), Finnland (FI), Kroatien (HR), Polen (PL); und 5) „quantitative Beurteilung, Druck auf Englisch zu publizieren“ mit Bosnien und Herzegowina (BA), Estland (EE), Ungarn (HU), Slowenien (SI), Slowakei (SK) (Ochsner et al., 2018). Die Evaluationssysteme vieler Länder zeichnen sich dabei aber dadurch aus, Mischtypen der verschiedenen Idealtypen zu sein.

16.3 Was ist Forschungsqualität, oder was ist der Gegenstand von Forschungsevaluation?

Evaluationen von Forschungsleistungen dienen dazu, etwas darüber auszusagen, wie „gut" Forschung ist. Was unter Forschungsqualität zu verstehen ist, bleibt meist lediglich implizit und lässt sich nur von der Auswahl an Indikatoren ableiten, die im Evaluationsverfahren verwendet werden. In Verfahren, die auf Peer Review basieren, wird die Auslegung von Forschungsqualität oft sogar den einzelnen Begutachtenden überlassen. Wenden wir uns zunächst den indikatorenbasierten Verfahren zu. Die Auswahl der Indikatoren richtet sich hierbei normalerweise an der Verfügbarkeit der Daten aus: Mit der Verfügbarkeit von Zitationsdaten wurden in den letzten Jahrzehnten viele bibliometrische Indikatoren entwickelt, die für die Beurteilung von Forschungsleistungen verwendet werden. Mit dem Aufkommen sozialer Medien in der Wissenschaft wurde noch eine weitere Datenquelle erschlossen, in der Hoffnung, die Schwächen der zitationsbasierten Indikatoren auszumerzen. Bald wurde aber festgestellt, dass auch diese sogenannten Altmetriken ähnliche Nachteile mit sich bringen: Es ist nicht klar, was denn überhaupt gemessen wird, denn sowohl den Zitationen wie auch den Tweets, Reads, Downloads etc., aus denen die Altmetriken berechnet werden, liegen sehr verschiedene Motivationen oder Ursachen zugrunde, wie zum Beispiel im Falle von Zitationen Kritik, Kontrastierung, Anschluss an die Literatur oder Tradition (Bornmann und Daniel, 2008b; Moed, 2005; Tahamtan und Bornmann, 2018), und bezüglich Altmetriken gesellschaftlicher Nutzen, Verwendung von Forschungsresultaten in der Praxis oder der Gesellschaft, Satire (Verunglimpfung), Wissenschaftsbetrug oder Verfügbarkeit des Artikels (Bornmann und Haunschild, 2018; Gumpenberger, Glänzel und Gorraiz, 2016; Lanamäki, Usman und Ochsner, 2019). Die Indikatoren sind also meist nur lose mit dem zu messenden Gegenstand, nämlich Forschungsqualität, verbunden, wie auch in den letzten Jahren in der bibliometrischen Community vermehrt festgestellt wurde (siehe z. B. Brooks, 2005; CWTS, 2012; Donovan, 2007). Allerdings würde die Methodenlehre verlangen, dass zuerst überlegt werden sollte, was gemessen werden will, bevor man Indikatoren bestimmt, nicht umgekehrt (Borsboom, Mellenberg und van Heerden, 2004; für die Forschungsevaluation, siehe Schmidt, 2005, S. 3).

Bezüglich Peer Review sieht es nicht viel besser aus. Während die formalen Probleme von Peer Review, wie Reliabilität, prädiktive Validität, verschiedene systematische Fehler wie Voreingenommenheit, Subjektivität usw. in der entsprechenden Literatur abgehandelt werden (siehe z. B. Bornmann und Daniel, 2008a; Daniel, Mittag und Bornmann, 2007; Langfeldt, 2010), interessiert hier das tieferliegende Verhältnis zwischen Forschungsqualität und Peer Review. Begutachtende wenden jeweils ihr eigenes Qualitätsverständnis an, das in der Regel implizit bleibt. Es bleibt nicht nur für andere implizit, was die Nachvollziehbarkeit und Vergleichbarkeit von verschiedenen Gutachten erschwert oder verunmöglicht, sondern auch für die begutachtende

Person selbst, denn die Urteile werden im Normalfall „holistisch“ gefällt, wie dem bekannten Beispiel aus Michèle Lamonts umfassender Studie zu Peer Review zu entnehmen ist: „Academic excellence? I know it when I see it“ (Lamont, 2009, S. 107; siehe auch Gozlan, 2016). Umfassende psychologische Studien zum Prozess der Entscheidungsfällung zeigen aber, dass sich solche holistischen Entscheidungen – in der Umgangssprache auch Bauchentscheide genannt – nicht eignen, um Leistungen oder Verdienste zu beurteilen (Thorngate, Dawes und Foddy, 2009, S. 21). In solchen holistischen Entscheidungen ändert die beurteilende Person die Qualitätsdefinition oder zumindest die Gewichtung verschiedener Aspekte potenziell für jedes Evaluationsobjekt, was Tür und Tor für eine ganze Palette systematischer Fehler, wie beispielsweise Bevorzugung von einigen Themen oder Methoden, Konservatismus, Bevorzugung von Autoren renommierter Institutionen, Geschlechterdiskriminierung usw., öffnet.

Die quantitativen wie auch die qualitativen Evaluationsverfahren leiden also offensichtlich unter demselben Problem: Es ist nicht klar, was denn genau gemessen oder beurteilt wird. Solange der Frage ausgewichen wird, was denn Forschungsqualität ist, weil es ohnehin unmöglich ist, eine allgemeingültige Definition dafür zu finden (Abramo und D’Angelo, 2016; Glänzel, Thijs und Debackere, 2016), kann sie auch nicht sinnvoll gemessen oder beurteilt werden. In anderen Worten: Es besteht ein Validitätsproblem. Unter Validität sei hier ganz einfach das Ausmaß verstanden, in welchem ein Maß (z. B. ein Indikator oder eine Beurteilung) das misst, was es zu messen verspricht (Kelley, 1927, S. 14). Das heißt, dass das zu messende Konstrukt bekannt sein muss und definiert werden muss, bevor es gemessen werden kann. Dies ist eine konzeptionelle Aufgabe, die weder mittels statistischer Methoden noch anhand von Daten erledigt werden kann (Borsboom et al., 2004, S. 1062). Forschungsqualität kann dabei als latentes Konstrukt aufgefasst werden: Latente Konstrukte sind Phänomene, die nicht direkt gemessen werden können, sondern etwas beschreiben, das messbaren Objekten zugrunde liegt. Solche latenten Konstrukte sind ubiquitär in der sozialwissenschaftlichen Forschung: Intelligenz, sozioökonomischer Status, Depression, soziale Vernetzung sind nur einige Beispiele. Entsprechend bietet die sozialwissenschaftliche Literatur auch Methoden für die Messung – oder in diesem Zusammenhang treffender: Operationalisierung – solcher latenten Konstrukte (Bollen, 2002). In unserem Fall ist das latente Konstrukt ein *a priori* Konstrukt, also etwas, wovon wir wissen, dass wir es messen wollen (im Gegensatz zu *a posteriori* Konstrukten, die sich aus den Daten ergeben und deren Ziel es primär ist, Variablen zusammenzufassen, um die Komplexität zu reduzieren, siehe Bollen, 2002). Will ein solches latentes Konstrukt gemessen werden, besteht der erste Schritt darin, die verschiedenen Dimensionen ausfindig zu machen, die das Konstrukt ausmachen (Borsboom et al., 2004). Für diese Dimensionen können dann Indikatoren gefunden werden oder – falls das Konstrukt sehr komplex ist – weitere Aspekte definiert werden, die mit Indikatoren messbar sind (vgl. bezüglich Forschungsqualität Hug, Ochsner und Daniel, 2014, S. 60).

Um ein solches latentes Konstrukt konzeptionell zu definieren, wird ein Katalog an Kriterien benötigt, der alle relevanten Aspekte von Forschungsqualität (oder For-

schungsleistung) abdeckt. Eine solche Konzeptualisierung des zu messenden Gegenstands hat den Vorteil, dass klar wird, was der zu messende Gegenstand genau ist. Es können den einzelnen Dimensionen und Aspekten Indikatoren zugewiesen werden, und so wird auch explizit, welche Dimensionen von Forschungsqualität mit Indikatoren erfasst werden können und welche nicht. Für die Beurteilung von Forschungsleistungen durch Peer Review ergibt sich mit einem solchen Kriterienkatalog ein Leitfaden für die Beurteilung, die somit angemessener wird, da die einzelnen Dimensionen bewertet werden und eine Gewichtung der verschiedenen Dimensionen explizit vorgenommen werden muss (Thorngate et al., 2009). Abbildung 16.2 zeigt eine schematische Darstellung einer solchen Konzeptualisierung.

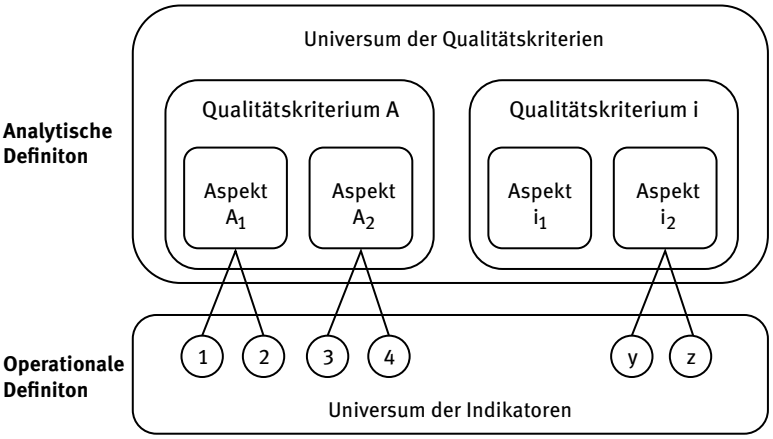

Abb. 16.2: Schematisches Messmodell für Forschungsqualität (Quelle: Hug, Ochsner und Daniel, 2014, S. 60).

16.4 Kriterien für Forschungsqualität in nationalen Evaluationskontexten

Kriterienkataloge für Forschungsqualität können in verschiedenen Kontexten gefunden werden, z. B. in Richtlinien für Gutachten für Zeitschriftenbeiträge oder in nationalen Evaluationsverfahren von Forschungsleistungen. Allerdings sind diese Kriterienkataloge meist sehr vage gehalten. In den letzten Jahren wurden jedoch vonseiten der Hochschulforschung, des Qualitätsmanagements an Hochschulen, aber auch vonseiten der Forschenden selbst einige Anstrengungen unternommen, einen breiteren Katalog an Kriterien zu erarbeiten. Einen vollständigen Überblick über alle Kriterienkataloge zu bieten ist nicht Aufgabe dieses Beitrags. Vielmehr geht es darum, verschiedene Ansätze aufzuzeigen und auf die Wichtigkeit hinzuweisen, Forschungsqualität

umfassend zu konzeptualisieren. Im Folgenden wird deshalb eine enge Auswahl der bekanntesten und relevantesten Kataloge vorgestellt.

In England wird seit 1986 das sogenannte Research Assessment Exercise (RAE), das später in Research Excellence Framework (REF) umbenannt wurde, durchgeführt. Es handelt sich dabei um eine umfassende Evaluation von Forschung an englischen Hochschulen. Die Resultate der Evaluation haben einen Einfluss auf die Finanzierung der Hochschulen. Die Evaluationsprozedur wandelt sich dabei fortlaufend (siehe z. B. Bence und Oppenheim, 2005; Moed, 2008). Im Unterschied zu den meisten sogenannten Performance-Based Research Funding Systems (PRFS) basieren die Evaluationen zu großen Anteilen auf Gutachten von Forschenden mit mehr oder weniger Einfluss von Forschungsindikatoren. Von besonderer Bedeutung für das Vorhaben dieses Beitrags ist das RAE 2008, das stärker auf eine Expertenevaluation der Qualität der Forschung setzte, ohne auf den Publikationskanal zu achten (Bence und Oppenheim, 2005). Dazu wurden drei Kriterien festgelegt, die stark an Polanyis (1962) theoretisch hergeleiteten Kriterien *Plausibility*, *Scientific Value* und *Originality* angelehnt sind: *Originality*, *Rigour* und *Significance*, also Originalität, Wissenschaftlichkeit und Bedeutsamkeit. Offensichtlich sind diese Kriterien sehr allgemein gehalten. So lag es in den Händen der Mitglieder der verschiedenen Panels zu bestimmen, was diese Kriterien in ihrem Fachbereich genau bedeuten. 15 Panels aufgeteilt in 67 Sub-Panels gegliedert nach Disziplinen waren im RAE 2008 damit beauftragt, die Qualität von Forschungsprodukten zu bestimmen. Hunderte von Forschenden nahmen so Einfluss darauf, wie Forschungsqualität definiert wird. Allerdings zeigt eine Analyse der verschiedenen Definitionen, dass auch die fachspezifischen Definitionen sehr vage waren und zumeist sehr ähnlich blieben, was den Autor einer systematischen Analyse dieser Definitionen zu folgender Schlussfolgerung bewog: „The words may be different but the imprecision is just the same so that, like pornography, academic excellence can be recognised but not defined“ (Johnston, 2008, S. 131).

In Deutschland entwickelte der Wissenschaftsrat 2004 als Reaktion auf die immer stärker an Bedeutung gewinnenden aber methodisch unzureichenden Wissenschaftsrankings das sogenannte „Forschungsrating“ (Wissenschaftsrat, 2004). Es sollte dazu dienen, „Universitäten und außeruniversitären Forschungseinrichtungen im Rahmen ihrer jeweiligen Mission und in Verbindung mit anderen Verfahren der Qualitätssicherung und strategischen Planung in ihren strategischen Entscheidungen und bei der Qualitätssicherung in der Forschung zu unterstützen und den Qualitätswettbewerb zu fördern“ (Wissenschaftsrat, 2004, S. 43). Dies soll mittels fachspezifischer Forschungsprofile geschehen, die durch die zu bewertenden Einheiten einzureichen sind. Eine Beurteilung findet dabei durch Bewertungsgruppen für jedes Fachgebiet einzeln statt und erfolgt multidimensional nach neun Rahmenkriterien in drei Dimensionen. Eine Gewichtung der Kriterien oder eine Verrechnung zu einer Gesamtnote sind explizit nicht vorgesehen (Wissenschaftsrat, 2004, S. iii–iv). Folgende Kriterien fließen in die Bewertung ein: Dimension *Forschung*: (1) Qualität, (2) Effektivität, (3) Effizienz; Dimension *Nachwuchsförderung*: (1) Prozesse der Nach-

wuchsförderung, (2) Erfolge der Nachwuchsförderung; Dimension *Wissenstransfer*: (1) Relevanz, (2) wirtschaftliche Umsetzung, (3) Fort- und Weiterbildung, (4) forschungsbasierte Beratung, Wissenschaftskommunikation (Wissenschaftsrat, 2004, S. 45–47). Aus diesem Kriterienkatalog wird sofort ersichtlich, dass die Beurteilung weit über die Beurteilung der Forschungsqualität hinausgeht. Forschungsqualität ist eines von neun Kriterien und es ist folgendermaßen spezifiziert: „Dieses Kriterium umfasst die Aktualität und Relevanz der Fragestellungen für das Forschungsgebiet, Neuheit und Originalität der Forschungsergebnisse sowie die Eignung und Verlässlichkeit der Methoden. Daten über die fachliche Resonanz der Forschungsergebnisse (z. B. normierte relative Zitationsindikatoren) können als Indikatoren genutzt werden" (Wissenschaftsrat, 2004, S. 45). Es zeigt sich, dass sich die Definition kaum von jener des RAE unterscheidet: Aktualität und Relevanz (Significance), Neuheit und Originalität (Originality), Eignung und Verlässlichkeit der Methoden (Rigour). Hinzu kommt noch die fachliche Resonanz, die aber als Indikator für alle drei Kriterien aufgefasst wird. Das Forschungsrating wurde in verschiedenen Fächern erprobt (Wissenschaftsrat, 2008a, 2008b, 2008c). Während es in Chemie und Elektro- und Informationstechnik grundsätzlich auf Akzeptanz stieß, war die Kritik in der Soziologie etwas deutlicher (siehe z. B. Riordan, Ganser und Wolbring, 2011). In den Geisteswissenschaften allerdings führte das Rating zu großen Kontroversen und der Verband der Historiker und Historikerinnen Deutschlands (VHD) boykottierte die Teilnahme (für eine Zusammenfassung der Kritik, siehe Ochsner, Hug und Daniel, 2012b; im Originalton ist die Kritik u. a. nachzulesen bei Plumpe, 2009, 2010; Hose, 2009). Darauf wurde eine Unterarbeitsgruppe von Geisteswissenschaftler(inne)n einberufen, um das Rating an die Geisteswissenschaften anzupassen. Die Unterarbeitsgruppe nahm Anpassungen an den Kriterien vor und empfahl, die Bewertung primär durch Peer Review durchzuführen und quantitative Indikatoren nur zur Ergänzung beizuziehen. Anschließend wurde das Forschungsrating für die Geisteswissenschaften an der Anglistik/Amerikanistik erprobt, deren Bewertungsgruppe das Verfahren nochmals angepasst hat, was zu folgendem Katalog mit nun vier Dimensionen führte (Wissenschaftsrat, 2011): Dimension *Forschungsqualität*: (1) Qualität des Outputs, (2) Quantität des Outputs; Dimension *Reputation*: (1) Anerkennung, (2) Professional Activities; Dimension *Forschungsermöglichung*: (1) Drittmittelaktivitäten, (2) Nachwuchsförderung (3) Infrastrukturen und Netzwerke; Dimension *Transfer an außeruniversitäre Adressaten*: (1) Personaltransfer, (2) Wissensvermittlung. Auffallend ist dabei insbesondere die neue Dimension „Reputation" sowie das neue Kriterium „Infrastrukturen und Netzwerke". Forschungsqualität ist dabei nicht genauer spezifiziert, es wird lediglich zwischen Qualität (exemplarische Publikationen und Publikationsliste) und Quantität des Outputs (Anzahl Publikationen) unterschieden (Wissenschaftsrat, 2011; für eine reflexive Analyse des Forschungsratings der Anglistik/Amerikanistik siehe z. B. Hornung, Khlavna und Korte, 2016; Plag, 2016).

Ein weiteres Beispiel für die Entwicklung von Qualitätskriterien im Rahmen einer nationalen Evaluation bietet das Standard Evaluation Protocol (SEP) der Niederlande. Sämtliche Forschungseinheiten werden alle sechs Jahre mittels des SEP evaluiert. Das SEP 2003–2009 sah folgende Kriterien vor: Qualität, Produktivität, Relevanz und Machbarkeit. Die Kriterien wurden dabei bewusst flexibel gehalten, damit die Begutachtenden eine fachspezifische Spezifikation vornehmen können. Die Gutachten sollten dabei sowohl quantitativ sein wie auch eine reflexive Komponente beinhalten (VSNU, VSO und KNAW, 2003, S. 16). In der Checkliste für die Begutachtenden werden die Kriterien allerdings deutlicher spezifiziert (VSNU et al., 2003, S. 40). Qualität wird dabei anhand der folgenden Aspekte beurteilt: (1) Originalität, (2) Bedeutsamkeit, (3) Kohärenz des Forschungsprogramms, (4) Publikationsstrategie, (5) Prominenz des Direktors oder der Direktorin, (6) Prominenz anderer Forschenden der Einheit, (7) Qualität der wissenschaftlichen Publikationen (wissenschaftlicher Impact), (8) Qualität anderer Resultate. Diese Spezifikation weist darauf hin, dass zwar eine qualitative Begutachtung im Sinne einer Reflexion der Interpretation quantitativer Daten Teil der Evaluation ist; die Ausführungen zu den Informationen, die der Begutachtung zugrunde liegen, zeigen aber, dass quantitative Indikatoren eine große Rolle spielen: So stehen wissenschaftliche Publikationen, Patente und die Höhe der eingeworbenen Drittmittel im Zentrum, während Publikationen und Aktivitäten, die nicht auf das einschlägige wissenschaftliche Publikum zielen, explizit exkludiert werden. Dies führte zu einiger Kritik insbesondere aus den Sozial- und Geisteswissenschaften und zu einer Reihe von Arbeitsgruppen und Reports, die den Zweck geisteswissenschaftlicher Forschung reflektieren und in Bezug zur Evaluation stellen (Committee for the National Plan for the Future of the Humanities, 2009; KNAW, 2005; 2009). Das schließlich eingesetzte „Committee on Quality Indicators in the Humanities" konstatierte, dass einerseits die Wissenschaftspolitik zu viel von Indikatoren erwarte, andererseits gerade in den Geisteswissenschaften eine zu große Aversion gegen „Messung" von Forschungsleistungen bestehe und schlug deshalb einen Mittelweg vor. Dieser sieht eine Evaluation mittels Informed Peer Review entlang zweier Dimensionen vor: wissenschaftlicher Output und gesellschaftliche Qualität. Jede der Dimensionen wird anhand dreier Kriterien beurteilt: (1) wissenschaftliche/gesellschaftliche Publikationen, (2) wissenschaftliche/gesellschaftliche Nutzung von Forschungsoutput und (3) Evidenz wissenschaftlicher/gesellschaftlicher Anerkennung. Alle drei – resp. sechs – Kriterien werden mit Unterstützung quantitativer Indikatoren qualitativ von Begutachtenden beurteilt (KNAW, 2011, S. 47). Diese Anpassung der Kriterien scheint auch die anderen Disziplinen überzeugt zu haben, so verläuft die Evaluation des SEP 2015–2021 nun ebenfalls entlang der zwei Dimensionen Forschungsqualität und gesellschaftliche Relevanz spezifiziert durch die drei oben genannten Kriterien (VSNU et al., 2016, S. 7, 25).

Diese drei Beispiele nationaler Evaluationsprozeduren zeigen, dass die Definition von Forschungsqualität Schwierigkeiten bereitet und insbesondere in den Geisteswissenschaften kritisiert wird. Allerdings zeigt gerade die Entwicklung des SEP, dass dies

nicht nur damit zu tun hat, dass die eingesetzten Kriterien nicht auf die Geisteswissenschaften passen, sondern auch damit, dass sich die Geisteswissenschaften stärker im Vorfeld damit befassen, was wozu gemessen werden soll und welche Effekte dies haben könnte, während die eher experimentell ausgerichteten Natur-, Lebens- und Technikwissenschaften dazu bereit sind, zuerst zu messen und anschließend Anpassungen vorzunehmen. Ersichtlich wird an diesen Beispielen auch, dass die Definitionen von Forschungsqualität in solchen Prozeduren sehr allgemein gehalten sind, da eine solche Top-down-Bestimmung der Kriterien für alle Fächer gültig sein muss und die Kriterien ohne vorhergehende Abklärung über Fächerunterschiede bezüglich Qualitätsdefinitionen eben auch vage bleiben müssen.

Anstatt Kriterien top-down vorzugeben, könnten Kriterien jedoch auch bottom-up aus den Fächern selbst heraus entwickelt werden und von der Frage ausgehen, wie denn Forschende selbst über Forschungsqualität denken. Können Qualitätskriterien identifiziert werden, die für viele Fächer gelten, aber dennoch dem Begriff etwas Leben einhauchen können? Forschungsqualität ist kaum abschließend zu definieren. Somit sind es Perzeptionen von Qualität, die es zu messen gilt. Von besonderem Interesse sind die Qualitätsvorstellungen der Forschenden selbst, denn diese können die Qualität von Forschung in ihrem Feld am besten beurteilen (Hemlin, 1993; LERU, 2012). Im Folgenden werden deshalb Studien vorgestellt, die Qualitätskriterien für die Forschung empirisch herleiten.

16.5 Empirisch hergeleitete Qualitätskriterien für Forschung

16.5.1 Die Anfänge oder die 1960er bis 1980er: was macht einen guten Text aus?

In den ersten empirischen Studien über Qualitätskriterien der Forschenden selbst interessierten sich die Autoren dafür, wie Normen für wissenschaftliche Forschung (z. B. Merton, 1957) in der Praxis verbreitet sind (z. B. Chase, 1970; Lindsey und Lindsey, 1978), für Untersuchungen, wie Herausgebergremien sowie Begutachtende Artikel beurteilen (sogenanntes Manuscript Reviewing, z. B. Kerr, Tolliver und Petree, 1978; Rowney und Zenisek, 1980; Wolff, 1970), oder wie man Studierenden oder Nachwuchsforschenden erklären kann, wie man einen guten Artikel schreibt (Frantz, 1968). Diese Studien haben gemeinsam, dass sie der Professorenschaft, Mitgliedern von Herausgebergremien und Begutachtenden Kriterien zur Bewertung vorlegten. Die Kriterien wurden von den Autor(inn)en der jeweiligen Studie zusammengestellt, meist anhand von Kriterien, die von Zeitschriften vorgeschlagen oder aus den Theorien wissenschaftlicher Normen abgeleitet wurden, und den Forschenden zur Bewertung vorgelegt, ohne dass diese selbst neue Kriterien nennen konnten. Während diese Studien also nicht direkt Einblick in die Qualitätsverständnisse der Forschenden geben, gewähren sie dennoch einen Blick darauf, wie die Forschenden einzelne vorgegebene Kriterien gewichten. Zusammenfassend lässt sich festhalten, dass in den Studien

Kriterien wie Wissenschaftlichkeit (Replizierbarkeit, Design, Klarheit des Textes) und Originalität (Beitrag zum Wissensstand, neue Theorie, neue Methode, Kreativität) ganz oben auf der Liste stehen, während sich die Eigenschaften der Person[66] (Status, Affiliation, Reputation) oder des Textes (Länge, Punktation, Unterhaltungswert) unten in der Prioritätenliste wiederfinden. Ebenfalls tiefe Bewertungen erhält fast durchgängig die Anwendbarkeit. Allerdings besteht bezüglich dieses Kriteriums Uneinigkeit, die sich durch eine deutlich größere Standardabweichung bemerkbar macht (Chase, 1970; Frantz, 1968; Lindsey und Lindsey, 1978; Wolff, 1970). Während sich die meisten dieser Studien einem Fach widmeten (Erziehungswissenschaften: Frantz, 1968; Psychologie: Rowney und Zenisek, 1980; Wolff, 1970), fanden Chase (1970) Unterschiede zwischen Natur- und Sozialwissenschaften und Lindsey und Lindsey (1978) zwischen Psychologie, Soziologie und Sozialarbeit. Die Unterschiede sind – wenig überraschend – insbesondere bezüglich der Gewichtung von Theorie und Empirie auszumachen. So berichtet Chase (1970), dass in den Sozialwissenschaften der Wissenschaftlichkeit, der theoretischen Bedeutung und der Anwendbarkeit statistisch signifikant mehr Bedeutung zugewiesen wird als in den Naturwissenschaften. Auf der anderen Seite ist in den Naturwissenschaften die Replizierbarkeit, mathematische Präzision, die Abdeckung der bestehenden Literatur und zu einem bescheidenen Grad auch die Originalität wichtiger als in den Sozialwissenschaften. Allerdings ist anzumerken, dass sich an der Rangfolge der Kriterien zwischen den Fächern nur wenig ändert. Lindsey und Lindsey (1978) gehen davon aus, dass es nicht nur zwischen Natur- und Sozialwissenschaften Unterschiede gibt, sondern dass sich solche Unterschiede auch zwischen Fächern der Sozialwissenschaften zeigen. Sie finden dann auch, dass in der Psychologie und der Soziologie den qualitativen Methoden weniger Relevanz zugewiesen wird als dies in der Sozialarbeit der Fall ist. Zur Überraschung der Autoren wird aber in der Psychologie den qualitativen Methoden noch ein wenig mehr Bedeutung zugemessen als in der Soziologie. Ihrer Interpretation nach könnte dies daran liegen, dass die Psychologie oft in Kritik stand, zu stark einem einzigen Paradigma zu folgen (Lindsey und Lindsey, 1978, S. 56).

16.5.2 Die 1990er und 2000er: Forschungsqualität im Rahmen von Evaluation

Den erwähnten Studien der frühen Periode ist ebenfalls gemeinsam, dass sie den Zeitschriftenartikel im Blick haben, also eine einzelne, klar abgrenzbare und im Umfang beschränkte Forschungsarbeit. Forschung geht aber insbesondere in den Sozial- und

66 Als einzige Ausnahme unter den Studien finden Rowney und Zenisek (1980), dass die Reputation der Autor(inn)en durchaus einen Einfluss auf die Akzeptanz eines Zeitschriftenartikels haben kann. Dies hat wohl mit dem unterschiedlichen Kontext zu tun (Beurteilung, welche Aussagen die Annahme eines Artikels beeinflussen; Formulierung „ich kenne den Autor und finde, dass er eine gerechtfertigte Reputation im thematischen Bereich genießt“).

Geisteswissenschaften über das Schreiben von Zeitschriftenartikeln hinaus. Ebenso geht Forschungsqualität darüber hinaus, was und wie etwas in einer Zeitschrift publiziert werden kann und soll. Hartmann und Neidhardt (1990) untersuchten beispielsweise, wie Forschende Kriterien in der Beurteilung von Forschungsanträgen der Psychologie, Politikwissenschaften, ökonomischer Theorie und Elektroingenieurswesen benutzen. Sie untersuchten, wie häufig die aus der Literatur hergeleiteten Kriterien in Beurteilungen vorkommen und welche Kriterien die Schlussbewertung am meisten beeinflussen. Theorie, Methode, Budget und wissenschaftliche Relevanz sowie etwas weniger stark die praktische Relevanz sind demnach wichtige Kriterien für die Beurteilung von Forschungsanträgen (einen ähnlichen Ansatz verfolgt auch Gulbrandsen, 2000).

Eine frühe systematischere Auseinandersetzung mit Qualitätsvorstellungen von Forschenden stellt die Studie von Sven Hemlin und Henry Montgomery aus dem Jahre 1990 dar (Hemlin und Montgomery, 1990). Mittels Interviews mit 22 schwedischen Forschenden mit Evaluationserfahrung prüfen sie, ob ihr Konzept von Forschungsqualität mit den Ideen der Forschenden und den Kriterien einiger der oben genannten Untersuchungen übereinstimmt. Ihr Konzept von Forschungsleistung umfasst dabei sieben Faktoren: Forschungspolitik (Research Policy), forschende Person, Forschungsleistung (Research Effort), Forschungsumfeld, Qualitätsindikatoren, innerwissenschaftliche Effekte und außerwissenschaftliche Effekte. Dabei unterscheiden sie zusätzlich zwischen Attributen (etwa „neu“, „präzise“) und Aspekten (etwa „Methode“, „Theorie“), allerdings ist nicht so klar, bei welchen Faktoren sie diese Unterscheidung anwenden. Sie finden, dass ihre Resultate mit denjenigen von Chase (1970), Frantz (1968), Rowney und Zenisek (1980) und Wolff (1970) übereinstimmen. Ihr Konzept von Forschungsqualität ist aber weitgreifender, denn die meisten Kriterien der früheren Studien fallen in die Faktoren Forschungsleistung (Research Effort) und forschende Person. Bemerkenswert an dieser Konzeption von Forschungsqualität ist, dass Qualitätsindikatoren einen eigenen Faktor bilden. Sie sind also nicht an Forschungsleistung, Person oder Umfeld geknüpft, sondern getrennt aufgeführt. Dies stellt die Frage, was denn die Funktion solcher Qualitätsindikatoren sein soll, wenn es sich dabei um einen getrennten Faktor handelt. Die Autoren schließen außerdem aus den Interviews, dass Forschende der Naturwissenschaften weniger gut als Forschende anderer Disziplinen beschreiben können, was denn für sie gute Forschung ist. Aufgrund der sehr kleinen Anzahl befragter Forschenden ist dieses Resultat aber mit Vorsicht zu genießen.

Mit der steigenden Relevanz von allgemeinen Forschungsevaluationen wird die Frage wichtig, welche Qualitätskriterien für Forschung in einem längerfristigen Kontext relevant werden und welche Kriterien und Konzeptionen von den bewertenden Forschenden angewendet werden. Hemlin (1993) untersucht mittels einer umfangreichen Befragung von 400 schwedischen Forschenden (224 Antworten; 56 Prozent Antwortrate), wie Forschende wissenschaftliche Arbeiten evaluieren. Als Basis verwendet er die oben vorgestellte Konzeption von Forschungsqualität (Hemlin und Montgome-

ry, 1990) mit einer kleinen Anpassung: Die Kriterien *inner-* und *außerwissenschaftliche Effekte* wurden zu einem Kriterium zusammengefasst und *Forschungsfinanzierung* als neues Kriterium hinzugefügt. Als Hauptresultat präsentiert Hemlin (1993) Fächerunterschiede in der Relevanz verschiedener Kriterien: Die Natur- und Technikwissenschaften auf der einen Seite und die Sozial- und insbesondere Geisteswissenschaften auf der anderen Seite betonen unterschiedliche Kriterien. Die Geisteswissenschaften zeichnen sich dadurch aus, dass sie den Schreibstil, die Argumentation, politische und kulturelle Effekte, Stringenz, Theorie, kreative Forschung sowie die Kreativität und Intelligenz der Person als deutlich wichtiger bewerten als die anderen Fächer, während sie das physische Forschungsumfeld, internationale Kontakte, außerwissenschaftliche Effekte, zielgerichtete Forschung (Directed Research), industrielle und sektorielle Finanzierung und Forschungsevaluation als weniger wichtig ansehen. Die Natur- und Technikwissenschaften bewerten genau umgekehrt. Die Sozialwissenschaften liegen näher bei den Geisteswissenschaften, liegen aber bezüglich Produktivität und internationaler Kontakte den Naturwissenschaften nahe und empfinden die Kreativität und Intelligenz der Person als weniger wichtig als Forschende der Geisteswissenschaften. Dafür zeichnen sie sich durch die höchste Bewertung der politischen und kulturellen Effekte aus. Die applizierten „harten" Wissenschaften wie Medizin und in gewissem Maße Technikwissenschaften zeichnen sich ihrerseits dadurch aus, dass sie – erwartungsgemäß – die außerwissenschaftlichen Effekte und Relevanz und nicht staatliche Finanzierung höher bewerten.

Dass sich die Qualitätsvorstellungen nicht nur zwischen den „harten" und den „weichen" Wissenschaften unterscheiden, zeigen Guetzkow, Lamont und Mallard (2004) in ihrer Analyse von sozial- und geisteswissenschaftlichen Gutachten für Forschungsanträge in den USA. Sie fokussieren dabei auf das in ihrem Material am häufigsten vorkommende Kriterium: Originalität. Während Forschende der Geisteswissenschaften eher Originalität in der Vorgehensweise und den Daten hervorheben, empfinden Forschende der Sozialwissenschaften eher die Originalität der Methoden als zentral. Zusätzlich extrahierten Guetzkow, Lamont und Mallard aus den Gutachten die Kriterien Klarheit, gesellschaftliche Relevanz, Interdisziplinarität, Machbarkeit, Wichtigkeit, Breite, Sorgfältigkeit, Nützlichkeit und „spannend". Leider wurden diese Kriterien in der Studie nicht weiter auf Fachunterschiede untersucht.

Einen anderen Ansatz verfolgt die Studie von Montada, Krampen und Burkhard (1999), die ebenfalls zum Ziel hat, Kriterien zu identifizieren, die im Rahmen einer Evaluation eingesetzt werden können. Das Argument ist, dass eine schlichte Postulierung der Gültigkeit von Evaluationskriterien durch Ministerien oder Evaluationsagenturen nicht gerade zur Akzeptanz beitragen würde. Vielmehr müssten die Kriterien konsensual mit der Fachgemeinschaft entwickelt werden. Die Autoren legten der gesamten Professorenschaft der Psychologie in Deutschland 117 Indikatoren zur Beurteilung bezüglich ihrer Eignung für eine Forschungsevaluation vor. 36 Prozent der Indikatoren wurden dabei als höchst positiv und 50 Prozent als positiv eingeschätzt. Lediglich 14 Prozent wurden als wenig positiv beurteilt. Allerdings stellten die Autoren

bei vielen Indikatoren eine große Varianz fest, was sie als problematisch für die Konsensfähigkeit der Indikatoren ansehen. Die Autoren schließen aus den Bewertungen der Indikatoren, dass gegenüber Evaluationsverfahren, die auf nur wenigen Indikatoren beruhen, Bedenken geäußert werden sollten, da die Anzahl an positiv bewerteten Indikatoren sehr groß ist.

16.5.3 Die 2010er: Qualitätsvorstellungen und empirische Messmodelle

Von der Kritik ausgehend, dass Forschungsleistung bislang lediglich dadurch definiert wird, welche Indikatoren in den verschiedenen Evaluationsprozeduren zum Einsatz kommen, während der Definition, was denn Forschungsleistung bedeutet, ausgewichen wird, entwickeln Bazeley (2010), Andersen (2013) und das Team um Hans-Dieter Daniel (Hug, Ochsner und Daniel, 2013, 2014; Ochsner, Hug und Daniel, 2012a, 2013, 2014) mittels Mixed-Methods-Ansätzen jeweils ein empiriegeleitetes theoriebasiertes Konzept für Forschungsleistung. Bazeley (2010) befragte dazu die Forschenden aller Disziplinen an drei australischen Universitäten. Acht „Attribute“ wurden den Forschenden vorgelegt, die diese in Bezug auf ihrer Meinung nach leistungsstarke Forscher/-innen interpretieren und erweitern mussten: Qualität, Fähigkeit, Produktivität, Anerkennung, Nutzen, Aktivität, Befriedigung und Umgänglichkeit. Anhand der Beschreibungen dieser Attribute wurde iterativ ein Modell für Forschungsleistung erarbeitet. Das Modell besteht aus sechs Dimensionen, vier davon beschreiben Forschungsaktivität (Engagement, Aufgabenorientierung, Forschungspraxis und intellektueller Prozess) und zwei Dimensionen beschreiben die Sichtbarmachung von Forschung (Dissemination und kollegiales Engagement). Dabei müssen von den ersten vier alle Dimensionen erfüllt sein, während von den zweiten zwei nur eine Dimension erfüllt sein muss, um als leistungsstark beurteilt werden zu können. Andersen (2013) geht dem Qualitätskonzept von medizinischer Forschung nach. Mittels qualitativer Interviews und einem quantitativen Fragebogen erarbeitet er ein auf den Qualitätsvorstellungen von dänischen Medizinern basierendes Konzept von Forschungsqualität mit dem Ziel, ein besseres Verständnis vom Zusammenhang zwischen Forschungsqualität und evaluativer Bibliometrie zu erhalten. Aus den 32 Aspekten von Forschungsqualität, die er aus den Interviews extrahiert hatte, ermittelte er zwölf Kriterien mittels einer Faktorenanalyse (Prestige der Zeitschrift, klinische Guidelines, Zitationsverhalten, Methodenkapitel, subjektive Qualität, grundlagen- und anwendungsorientierte Forschung, Autor, Bedeutung von Zitation, Verhältnis von Qualität und Zitation, Innovation, Skeptizismus, Sorgfältigkeit). Die zwölf Kriterien machen dabei drei Dimensionen von Forschungsqualität aus: Dissemination, Effekte auf die Politik, Effekte auf die Gesundheit. Es fällt bei den Kriterien auf, dass – wie in der Medizin allgemein üblich – die Bibliometrie und die Anwendungsorientierung eine große Rolle spielen. Andersen schlägt schließlich ein Messmodell ähnlich einem nomologischen Netzwerk vor, das auf den drei Dimensio-

nen beruht. Die zwölf Kriterien sind dabei teilweise als zu messende Aspekte wiederzuerkennen. Die Quantifizierung des Modells lag aber nicht im Rahmen der Arbeit.

Im Projekt „Entwicklung und Erprobung von Qualitätskriterien für die Forschung in den Geisteswissenschaften am Beispiel der Literaturwissenschaften und der Kunstgeschichte“ gingen wir einen Schritt weiter und setzten den im Projekt entwickelten und im Abschnitt „Was ist Forschungsqualität“ erwähnten Messansatz um (Ochsner et al., 2014). Da das Wissen über Forschungsqualität primär implizit vorliegt – wie die bekannten Zitate zeigen, dass man gute Forschung schon erkenne, wenn man sie sehe (wie etwa in Gozlan, 2016, S. 271; Johnston, 2008, S. 131 oder Lamont, 2009, S. 107 dokumentiert) –, wurde in diesem Projekt „auf der grünen Wiese“ begonnen und nicht wie in den meisten der bisher erwähnten Projekten auf einem bereits aus der Literatur abgeleiteten Kriterienkatalog aufgebaut. Mittels einer speziellen Methode zur Explikation impliziten Wissens (Repertory Grid Interviews) wurde zuerst das Qualitätsverständnis von Geisteswissenschaftler(inne)n untersucht (Ochsner et al., 2013). Dabei zeigte sich, dass die Forschenden aller drei untersuchten Fächer, Deutsche Literaturwissenschaft, Englische Literaturwissenschaft und Kunstgeschichte, zwischen zwei Konzeptionen von Forschung unterscheiden: dem modernen und dem traditionellen Konzept von Forschung. Diese Konzepte sind unabhängig von der Qualität der Forschung. Somit ergeben sich vier Typen von Forschung: Die positiv konnotierte moderne Forschung umfasst dabei internationale, interdisziplinäre und kollaborative Forschung mit direkter gesellschaftlicher Relevanz, die negativ konnotierte moderne Forschung hingegen zeichnet sich durch Karriereorientierung aus, und Interdisziplinarität findet z. B. nur auf dem Papier statt, um Mittel akquirieren zu können. Die positiv konnotierte traditionelle Forschung repräsentiert das Genie. Sie beinhaltet exzellente Einzelforschung, die Potenzial zum Paradigmenwechsel hat, während die negativ konnotierte traditionelle Forschung dadurch gekennzeichnet ist, dass die Forschung isoliert und im Elfenbeinturm stattfindet. Dies zeigt, dass viele häufig eingesetzte Forschungsindikatoren nicht Forschungsqualität messen, sondern Indikatoren für das moderne Konzept von Forschung sind, das sowohl gute wie schlechtere Forschung beinhalten kann: Interdisziplinarität, Kollaboration, gesellschaftliche Relevanz und Internationalität sind demnach keine (oder zumindest nicht in jedem Fall) Qualitätsindikatoren. Auf diesen Erkenntnissen aufbauend wurde in einer mehrstufigen Delphi-Befragung von Forschenden der gleichen drei Fächer an den Schweizer Universitäten und an den Universitäten, die Mitglied der League of European Research Universities (LERU) sind, ein international abgestützter Qualitätskatalog entwickelt, der aus 19 Kriterien besteht, die durch 70 Aspekte spezifiziert sind (Hug et al., 2013; siehe zusammenfassend in Tabelle 16.1). Der Katalog wurde in anderen Projekten auch für andere Fächer (Rechtswissenschaften: Lienhard, Tanquerel, Flückiger, Amschwand, Byland und Herrmann, 2016; Schmied, Byland und Lienhard, 2018; Sozialwissenschaften: Ochsner und Dokmanović, 2017; Theologie: Mertens und Schatz, 2016) angepasst. Es zeigte sich, dass Fächerunterschiede sich meist lediglich in unterschiedlichen Gewichtungen der Kriterien zeigten, nur wenige Kriterien muss-

ten jeweils an die Fächer angepasst werden. Zum Beispiel wird die gesellschaftliche Relevanz (zu unterscheiden von Einfluss auf die Gesellschaft) in den Sozialwissenschaften höher gewichtet als in den Geisteswissenschaften. Der Einfluss auf die Praxis hingegen ist in den Rechtswissenschaften ein wichtiger Punkt. Im Gegensatz dazu ist die Pflege des kulturellen Gedächtnisses in den Geisteswissenschaften wichtiger als in den anderen Fächern. Der Katalog lässt sich außerdem an verschiedene Evaluationssituationen anpassen. Ochsner, Hug und Daniel (2017) adaptierten den Katalog für die Beurteilung von Forschungsanträgen von Nachwuchswissenschaftler(inne)n in den Geisteswissenschaften. Dabei wurden einige Kriterien, die nicht auf die Situation zutreffen, weggelassen (z. B. Offenheit für andere Personen) und Kriterien hinzugefügt (Machbarkeit, Person). Bei der Bewertung zeigten sich auch Unterschiede zur Bewertung der Kriterien für eine allgemeine Evaluation von Forschungsleistungen: So ist der Aspekt „ein neues Paradigma erschließen" des Kriteriums „Originalität" in der allgemeinen Evaluation von Forschungsleistungen eines der wichtigsten Items, während es in der Beurteilung von Anträgen von Nachwuchsforschenden zwar eine positive, aber vergleichsweise tiefe Bewertung erhält –, was dem Kontext Rechnung trägt, denn von einem Nachwuchsforschenden kann nicht erwartet werden, dass er mit einem relativ schwach dotierten Forschungsprojekt ein neues Paradigma generiert.

Aus diesen Studien lässt sich schließen, dass Forschungsqualität ein komplexes Konstrukt ist und viele Kriterien berücksichtigt werden sollten, will Forschungsqualität angemessen beurteilt werden.

Bisher lag der Fokus darauf, was denn Qualität ausmacht und welche Kriterien für die Beurteilung zu berücksichtigen sind. Anschließend stellt sich die Frage, ob diese Kriterien der Messung offenstehen: Welche Kriterien werden durch die häufig eingesetzten Indikatoren gemessen und welche nicht? Für welche Kriterien lassen sich Indikatoren finden und welche können nur durch Beurteilung durch Begutachtende bewertet werden? Im nächsten Abschnitt wird dementsprechend die Messbarkeit von Forschungsqualität erörtert.

16.6 Messbarkeit von Forschungsqualität am Beispiel der Geisteswissenschaften

Wie im Abschnitt „Was ist Forschungsqualität" erläutert, setzt eine Messung eine Spezifizierung des zu messenden Konzepts voraus. Besteht ein ausführlicher Kriterienkatalog – wie beispielsweise im vorhergehenden Kapitel vorgeschlagen –, kann mit diesem Wissen eruiert werden, welche Kriterien und Aspekte von Forschungsqualität mit den gängigen Indikatoren gemessen werden und welche Kriterien und Aspekte überhaupt der Messung offenstehen. In einer umfassenden Sammlung von Forschungsindikatoren auf Basis einer Vielzahl wissenschaftlicher Publikationen, Eva-

luationsverfahren sowie grauer Literatur wurde eine so große Anzahl an Indikatoren identifiziert, sodass sie in 62 Indikatorengruppen zusammengefasst wurden. Diese Indikatorengruppen wurden sodann Kriterien und Aspekten zugeteilt, die sie potenziell messen können (Ochsner et al., 2012a). Dies zeigte, dass nur rund 50 Prozent der Aspekte, die sich unter den Forschenden der Geisteswissenschaften als konsensual erwiesen haben, potenziell durch Indikatoren gemessen werden können. 50 Prozent der Aspekte sind hingegen nur in der Beurteilung durch Begutachtende bewertbar. Drastischer sieht es bezüglich der Indikatoren aus, die häufig in Evaluationsverfahren eingesetzt werden, wie zum Beispiel Zitationen, Preise, Drittmittel, Kollaborationen, Transfer in die Wirtschaft und Gesellschaft, Publikationen oder Expertenmandate. Diese messen mit Ausnahme zweier Kriterien (wissenschaftlicher Austausch und Einfluss auf die Wissenschaft) primär Aspekte gerade jener Kriterien, die unter den Forschenden schlecht bewertet wurden oder zumindest nicht konsensfähig waren, nämlich Produktivität, Reputation, Kontinuität, Wirkung auf die Gesellschaft und Relevanz (siehe Tabelle 16.1, Kriterien gekennzeichnet durch ††). Dies weist auf eine Diskrepanz zwischen den Qualitätsvorstellungen der Forschenden und der Evaluierenden hin.

Im Anschluss wurden die Forschenden gefragt, ob die Indikatoren ihrer Meinung nach die Aspekte, denen sie zugeordnet werden können, sinnvoll messen könnten. Hier fällt das Resultat nochmals deutlich negativer aus: Es besteht nur für sehr wenige Indikatoren Konsens bezüglich der Eignung für einen adäquaten Einsatz in der For-

Tab. 16.1: Qualitätskriterien und Messbarkeit durch Indikatoren (Quelle: Ochsner, Hug und Daniel, 2012a, siehe S. 4).

Wissenschaftlicher Austausch $^{**, ††}$	Kontinuität, Fortführung ††	Gelehrsamkeit, Belesenheit $^{**, †}$
Innovation, Originalität **	Wirkung auf die akademische Gemeinschaft $^{**, ††}$	Leidenschaft, Enthusiasmus $^{*, †}$
Produktivität ††	Gesellschaftsbezug, Wirkung auf die Gesellschaft ††	Konnex zwischen Forschung und Lehre, Scholarship of Teaching $^{**, †}$
Wissenschaftlichkeit **	Forschungsvielfalt $^{*, †}$	Forschungsvision $^{**, †}$
Pflege des kulturellen Gedächtnisses $^{**, †}$	Anschlussfähigkeit, Aktualität $^{**, †}$	Relevanz, Wichtigkeit ††
Reputation ††	Offenheit gegenüber Ideen und Personen $^{**, †}$	
Reflexion, Kritik $^{*, †}$	Selbststeuerung, Unabhängigkeit $^{*, †}$	

Bemerkungen: ** Kriterien, die in allen drei untersuchten Fächern Konsens erreicht haben,

* Kriterien, die in zwei der untersuchten Fächern Konsens erreicht haben;

†† Kriterium ist mit häufig eingesetzten potenziellen Indikatoren messbar,

† Kriterium ist potenziell mit Indikatoren messbar. Kriterien sind durch 70 Aspekte definiert, Messbarkeit bezieht sich auf mindestens einen Aspekt eines Kriteriums ist potenziell messbar.

schungsevaluation. Primär gilt dies für Indikatoren, die Forschungsaktivitäten aufzeigen, wie Anzahl Publikationen, Präsentationen oder andere Produkte, und die entsprechende Kriterien messen, wie wissenschaftlicher Austausch oder Pflege des kulturellen Gedächtnisses (Ochsner et al., 2014). Es zeigt sich, dass die Forschenden in den Geisteswissenschaften der Diskussion über Forschungsqualität gegenüber sehr offen sind, jedoch bezüglich der Messung eher skeptisch sind. Vielmehr ist denkbar, die Indikatoren als zusätzliche Information für eine qualitativ orientierte Beurteilung von Forschungsleistungen beizuziehen, denn die Studien zeigten, dass die Forschenden eine metrisch orientierte Evaluation als kritisch wahrnehmen, weil eine Reduktion auf messbare Aspekte der Komplexität von Forschungsleistung nicht gerecht werden kann und somit mit Risiken von Fehlanreizen verbunden ist.

Zwar sind, wie bereits erwähnt, die Geisteswissenschaften grundsätzlich der Messung weniger aufgeschlossen als die Naturwissenschaften. Nichtsdestotrotz zeigt sich an diesem Beispiel, wieso einer (rein) quantitativ orientierten Beurteilung von Forschungsleistung mit Vorsicht begegnet werden sollte. Forschungsqualität ist ein komplexes mehrdimensionales Konstrukt. Wenn in Evaluationen gewisse Aspekte systematisch nicht berücksichtigt werden, da sie nicht mit Indikatoren gemessen werden können, sind negative Effekte auf die Forschungspraxis nicht nur nicht auszuschließen, sondern sogar sehr wahrscheinlich. Solche negativen Effekte sind mittlerweile auch für Disziplinen des ganzen akademischen Spektrums gut belegt (siehe etwa Chavalarias, 2016; de Rijcke, Wouters, Rushforth, Franssen und Hammarfelt, 2016; Edwards und Roy, 2017; Kwok, 2013; Sousa und Brennan, 2014).

Was helfen nun diese Erkenntnisse im Hinblick auf eine adäquate Forschungsevaluation? Was tun, wenn sowohl Peer-Review-Verfahren problematisch als auch quantitative, vermeintlich objektive Methoden wenig erfolgversprechend sind? Im Folgenden wird ein Ansatz skizziert, der auf der Mehrdimensionalität des Konstrukts Forschungsqualität aufbaut und sowohl qualitative wie auch quantitative Formen der Beurteilung zulässt. Gleichzeitig vermag er als Bottom-Up-Ansatz die Qualitätsverständnisse verschiedener Anspruchsgruppen einzubinden, ohne sie zu vermengen und somit zu verdecken.

16.7 Bottom-up-Ansatz zur Beurteilung von Forschungsqualität

In vielen Situationen der Evaluation von Forschungsleistungen sind verschiedene Anspruchsgruppen involviert. Dies zeigt sich darin, dass häufig eingesetzte Forschungsindikatoren, wie beispielsweise Interdisziplinarität, Verbundforschung, Internationalität (häufig definiert als englischsprachig) oder Einfluss auf die Gesellschaft, aus Sicht der Forschenden nicht zwischen besserer und weniger guten Forschung zu unterscheiden vermögen, sondern lediglich als Indikatoren für eine neue Art von Forschung, nämlich der politisch gesteuerten, der Wissensgesellschaft verpflichteten Forschung, dienen können (siehe exemplarisch Rolfe, 2013, S. 3–20; und empirisch

Ochsner et al., 2013, S. 86). Daraus lässt sich für eine adäquate und transparente Evaluation schließen, dass Kriterien von verschiedenen Anspruchsgruppen separat ausgewiesen werden sollten, denn eine Vermischung von Kriterien von verschiedenen Anspruchsgruppen führt zu Kommunikationsproblemen und zu tieferer Akzeptanz bei den verschiedenen Anspruchsgruppen. Eine solche Vermischung kann auch zu systematischen Verzerrungen in einem Peer-Review-Verfahren führen, z. B. wenn Evaluierende interdisziplinäre Forschung evaluieren müssen, dies aber mittels ihrer jeweils eigenen disziplinären Standards tun (siehe Langfeldt, 2006).

Einfluss auf die Gesellschaft („Societal Impact") ist dabei eine eigene Dimension, denn dieser zielt nicht auf Forschungsqualität ab, sondern auf ein anderes Ziel der Forschung, nämlich deren Nutzen und Anwendbarkeit. Entsprechend sollte diese Dimension separat anhand ihrer eigenen Kriterien evaluiert werden (siehe auch KNAW, 2011; VSNU et al., 2016). Sogar die Begutachtenden für eine Beurteilung dieser Dimension können sich unterscheiden und neben Forschenden auch Personen aus der Zivilgesellschaft, der Wirtschaft oder der Politik umfassen. Dabei ist zu berücksichtigen, dass noch weniger Wissen darüber besteht, wie man denn den Einfluss auf die Gesellschaft beurteilen kann oder wie diese Dimension überhaupt zu definieren ist. Forschende bevorzugen die Beurteilung von Forschungsqualität vor der Beurteilung des Einflusses oder Nutzens der Forschung (Albert, Laberge und McGuire, 2012), und wenn sie letzteres doch tun müssen, sind sie unsicher dabei, was sie nun genau beurteilen sollen (Derrick und Samuel, 2017).

Eine rein quantitative Beurteilung von Forschungsleistung kann kaum zielführend sein, denn das Risiko von nicht intendierten negativen Effekten ist zu groß. Wenn weniger als 50 Prozent der relevanten Kriterien überhaupt mit Indikatoren messbar sind (siehe für die Geisteswissenschaften Ochsner et al., 2012a), wird klar, dass zu viele Aspekte von Forschungsqualität unter den Tisch fallen, damit eine solche Evaluation noch als adäquat angesehen werden kann. Während es sein kann, dass in den Natur- und Lebenswissenschaften mehr Kriterien als relevant angesehen werden, die effektiv messbar sind, ist es doch unwahrscheinlich, dass eine rein quantitative Evaluation alle wichtigen Aspekte von Forschungsqualität abdeckt. Insbesondere die generischen Kriterien Originalität und Wissenschaftlichkeit gehören gerade zu jenen Kriterien, die nicht durch Indikatoren abgebildet werden können.

Auch das Peer-Review-Verfahren ist harscher Kritik ausgesetzt. Es sei subjektiv, die Übereinstimmung der verschiedenen Begutachtenden sei tief, es führe zu konservativen Entscheiden, habe eine tiefe prädiktive Validität und sei von verschiedenen Verzerrungen betroffen, wie z. B. geschlechtsspezifischer Diskriminierung, Bevorzugung von großen Institutionen, Mainstreaming usw. (vgl. Bornmann und Daniel, 2008a; Bornmann, Mutz und Daniel, 2008; 2010; Mutz, Bornmann und Daniel, 2014; Tamblyn, Girard, Qian und Hanley, 2018). Allerdings liegt der Fokus der Kritik häufig auf dem Resultat des Peer-Review Prozesses, was mit verschiedenen methodologischen Problemen verbunden ist (Langfeldt, Bloch und Sivertsen, 2015): Es werden Resultate verglichen, ohne ein klares Konzept davon zu haben, was das Resultat denn

sein soll. So ist es zum Beispiel unklar, ob denn eine hohe Übereinstimmung der beurteilenden Personen überhaupt erwünscht ist. Eine Übereinstimmung könnte auch einfach nur eine Folge von einer unvorteilhaften Auswahl der Begutachtenden sein, die alle demselben Paradigma verpflichtet sind und deshalb jede Arbeit ablehnen, die nicht diesem Paradigma folgt. Ähnlich verhält es sich mit der prädiktiven Validität, die meist mit der Anzahl an Zitationen operationalisiert wird: Eine hohe prädiktive Validität kann auch nur von einer selbsterfüllenden Prophezeiung herrühren. Personen, die eine prestigeträchtige Finanzierung erhalten haben, erhalten mehr Aufmerksamkeit und dann mehr Zitationen, gerade weil sie die Finanzierung erhalten haben, nicht weil die Arbeit besser ist. Außerdem sind Zitationen von vielen Faktoren abhängig, nicht nur von der Qualität der Arbeit, und sie eignen sich deshalb nicht, Forschungsqualität zu messen (Bornmann und Daniel, 2008b; Moed, 2005; Ochsner et al., 2012a; Tahamtan und Bornmann, 2018). Schließlich können die Verzerrungen, die in Peer-Review-Verfahren ausfindig gemacht werden, auf andere Gründe als Unzulänglichkeiten im Peer-Review-Verfahren zurückzuführen sein: So könnten beispielsweise Forschende an prestigeträchtigen Institutionen oder Männer schlicht und einfach mehr Zeit und Ressourcen für Forschung haben als Forschende an kleinen, lehrorientierten Institutionen oder Frauen; oder letztere reichen weniger selbstbewusste Anträge ein als erstere. Beide Situationen würden dazu führen, dass Forschende an prestigeträchtigen Institutionen oder Männer bei eigentlich gleicher Eignung häufiger positiv beurteilt werden als Forschende an kleineren Institutionen oder Frauen, obwohl die Beurteilenden weder Männer noch prestigeträchtige Institutionen bevorzugen würden (siehe z. B. Ceci und Williams, 2011; Enserink, 2015).

Wichtiger für ein gelingendes Peer-Review-Verfahren ist die intra-rater Reliability (Ochsner et al., 2017), nämlich die Wahrscheinlichkeit, dass dieselbe bewertende Person für dasselbe Objekt zu verschiedenen Zeitpunkten dieselbe Bewertung abgibt, beispielsweise unabhängig von der Reihenfolge der zu bewertenden Objekte. In anderen Worten bedeutet dies, dass alle Objekte anhand derselben Standards beurteilt werden sollten. Um dies zu erreichen, sollte nach Thorngates' et al. (2009) Erkenntnissen zur Entscheidungsfindung und Beurteilung von Leistung jede Bewertung anhand einer Liste von allen relevanten Kriterien erfolgen. Dabei sollten die Kriterien separat bewertet werden, da „holistische" Urteile („ich erkenne Qualität, wenn ich sie sehe") sich dadurch auszeichnen, dass für jede Bewertung andere Gewichtungen der einzelnen Kriterien vorgenommen werden, was Tür und Tor für verschiedene Verzerrungen öffnet (Thorngate et al., 2009, S. 26). Außerdem verhindert die separate Bewertung einer Vielzahl an Kriterien, dass eine Bewertung durch Kriterien, die ähnlich bewertet werden, dominiert, denn Menschen tendieren dazu, Konsistenz in der Beurteilung zu bevorzugen anstatt Information zu maximieren, was zu schlechteren Entscheiden führt, wie Tversky und Kahnemann (1974, S. 1126) in ihrem grundlegenden Artikel zeigen. Auf Forschungsevaluation bezogen ist es bei holistischen Beurteilungen wahrscheinlich, dass ein Beitrag als „gut" bewertet wird, der bei Einzelbewertung der Kriterien vielleicht als problematisch angesehen worden wäre, beispielsweise, wenn die Kriteri-

en „Stil“ und „Relevanz“ als gut bewertet werden, aber „Nachvollziehbarkeit“ als ungenügend. Somit würden die gut bewerteten Aspekte „Stil“ und „Relevanz“ die holistische Beurteilung dominieren, während die Probleme in der Logik wenig beachtet würden, vielleicht gerade weil der „Stil“ die Argumentationsschwäche verdecken könnte. Schließlich führt eine Beurteilung anhand eines breiten Kriterienkatalogs auch zur Transparenz der Beurteilung: Anhand welcher Kriterien wurde beurteilt und wie wurden sie gewichtet? Dies dient außerdem den Autoren oder den Antragstellenden als Feedback und kann ihnen helfen, das nächste Mal einen besseren Text oder Antrag abzuliefern. Alle diese Punkte sind bedeutend für eine faire und konsistente Beurteilung von Forschungsleistung (Thorngate et al., 2009) und zeigen die Wichtigkeit einer klaren Konzeptualisierung von Forschungsleistung. Denn einem aus den Qualitätsvorstellungen der Fachgemeinschaft und den Anspruchsgruppen entwickelten umfassenden und expliziten Kriterienkatalog können also verschiedene Probleme von Peer-Review-Verfahren verhindert werden: Verzerrungen bezüglich des Geschlechts, Institutionen oder konservative Beurteilungen werden eher aufgedeckt werden; auch eher technische Probleme können reduziert werden, indem die separate Bewertung der Kriterien dabei hilft, zwischen Unterschieden der effektiven Beurteilung und zwischen unterschiedlichen Gewichtungen der einzelnen Kriterien zu unterscheiden, oder indem anhand der kriterienbasierten Beurteilung klar wird, wieso ein Antrag gefördert wurde, auch wenn die daraus stammenden Artikel vielleicht weniger oft zitiert werden, als die der nicht geförderten Antragssteller, weil Zitate vielleicht kein guter Gradmesser für den Erfolg des Projekts darstellen (z. B. weil ein noch nicht stark beforschtes Thema, oder eine auf Anwendbarkeit fokussierte Arbeit weniger Zitate generiert als ein Beitrag zu einem Mainstream-Thema).

Zusammenfassend kann für eine adäquate Beurteilung von Forschungsleistung auf die im Abschnitt „Was ist Forschungsqualität und wie kann sie gemessen werden?“ vorgestellte Konzeptualisierung von Forschungsqualität zurückgegriffen werden (Abbildung 16.2). Dabei sollten die Kriterien Forschungsqualität dem Fach und dem Kontext entsprechend definiert werden und Kriterien der verschiedenen Anspruchsgruppen einzeln ausgewiesen werden (siehe für eine solche Anpassung Ochsner et al., 2017). Experten beurteilen dann die Objekte der Evaluation separat für jedes Kriterium. Dabei können Kriterien auch Indikatoren zugewiesen werden, welche die Beurteilenden als Hilfe zur Bewertung beiziehen können (Informed Peer Review). Anschließend wird eine Gewichtung der Kriterien bestimmt, die für alle Objekte der Evaluation gleich ist.

16.8 Schlussfolgerungen

Die Beurteilung von Forschungsleistung ist eine komplexe Aufgabe. Bislang wurden die gängigen Verfahren als unzulänglich kritisiert. Quantitative indikatorenbasierte Verfahren werden kritisiert, Forschungsqualität nicht adäquat abzubilden und des-

halb unerwünschte Effekte zu provozieren, während qualitative Peer-Review-Verfahren kritisiert werden, subjektiv und mit vielzähligen Verzerrungen verbunden zu sein. In diesem Kapitel wird argumentiert, dass die Unzulänglichkeiten der gängigen Verfahren ihren Grund darin finden, dass das Konzept, das mit den Verfahren gemessen werden soll, nämlich Forschungsqualität, gar nicht erst explizit definiert wird und die Verfahren es deshalb auch nicht abbilden können.

Es wurden verschiedene Kriterienkataloge vorgestellt, die versuchen, Forschungsqualität fassbar zu machen. Es zeigte sich, dass Forschungsqualität ein komplexes, mehrdimensionales Konzept ist, das kontextabhängig definiert werden sollte. Schließlich wurde ein Ansatz skizziert, Forschungsleistung zu evaluieren, der auf der Mehrdimensionalität des Konstrukts Forschungsqualität aufbaut und sowohl qualitative wie auch quantitative Formen der Beurteilung zulässt. Gleichzeitig vermag er die Qualitätsverständnisse verschiedener Anspruchsgruppen einzubinden, ohne sie zu vermengen und somit zu verdecken.

Ausgangspunkt jeglicher Evaluation von Forschungsleistung sollte das Qualitätsverständnis der Forschenden im entsprechenden Fach sein, denn nur diese können die Qualität der Forschung wirklich beurteilen. Dies führt zu einer höheren Akzeptanz in der Forschungsgemeinschaft und hilft auch, unerwünschte Effekte zu verhindern. Die Kriterien sollten demnach bottom-up in der jeweiligen Disziplin erarbeitet werden. Diese Kriterien können dann durch Kriterien, die für andere Anspruchsgruppen relevant sind, ergänzt werden. Begutachtende sollen dann jedes Kriterium einzeln bewerten. Für alle zu bewertenden Objekte wird dieselbe Gewichtung der Kriterien verwendet. Falls möglich, können Indikatoren einzelnen Kriterien, die diese zu messen vermögen, zugewiesen werden, um die Begutachtenden bei der Beurteilung des Kriteriums zu unterstützen.

Dieser Ansatz wurde anhand der Geistes- und Sozialwissenschaften entwickelt (Hug et al., 2013; Ochsner und Dokmanović, 2017) und ist für verschiedene Evaluationssituationen adaptierbar (Ochsner et al., 2017). Während umfassende Kriterienkataloge für die Sozial- und Geisteswissenschaften bestehen, ist für die meisten natur- und lebenswissenschaftlichen Disziplinen noch ein entsprechender Kriterienkatalog zu erarbeiten (siehe auch Hug und Aeschbach, 2020). Einige Projekte haben sich diesem Ziel auch schon angenommen, so zum Beispiel das internationale vom norwegischen Research Council geförderte Kollaborationsprojekt R-Quest (www.r-quest.no).

16.9 Literaturverzeichnis

Abramo, G. und D'Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics*, 10(2):646–651. doi:10.1016/j.joi.2016.04.006.

Albert, M., Laberge, S. und McGuire, W. (2012). Criteria for assessing quality in academic research: the views of biomedical scientists, clinical scientists and social scientists. *Higher Education*, 64(5):661–676. doi:10.1007/s10734-012-9519-2.

Andersen, J. P. (2013). Conceptualising research quality in medicine for evaluative bibliometrics. Det Humanistiske Fakultet, Kobenhavns Universitet.

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education*, 35(8):889–903. doi:10.1080/03075070903348404.

Bence, V. und Oppenheim, C. (2005). The Evolution of the UK's Research Assessment Exercise: Publications, Performance and Perceptions. *Journal of Educational Administration and History*, 37(2):137–155. doi:10.1080/00220620500211189.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53(1):605–634. doi:10.1146/annurev.psych.53.100901.135239.

Bornmann, L., Mutz, R. und Daniel, H.-D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS One*, 5(12), e14331. doi:10.1371/journal.pone.0014331.

Bornmann, L. und Daniel, H.-D. (2008). Die Effektivität des Peer-Review-Verfahrens: Übereinstimmungsreliabilität und Vorhersagevalidität der Manuskriptbegutachtung bei der Angewandten Chemie. *Angewandte Chemie*, 120(38):7285–7290.

Bornmann, L., Mutz, R. und Daniel, H.-D. (2008). How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics*, 2(4):280–287.

Bornmann, L. und Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80.

Borsboom, D., Mellenbergh, G. J. und van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4):1061–1071. doi:10.1037/0033-295X.111.4.1061.

Bozkurt Umur, I., Diaz-Bone, R. und Surdez, M. (2017). How to evaluate research and teaching in sociology? Results of the survey conducted with members of Swiss Sociological Association (SSA). Report prepared for the Swiss Sociological Association. Swiss Sociological Association.

Brooks, R. (2005). Measuring University Quality. *The Review of Higher Education*, 29(1):1–21. doi:10.1353/rhe.2005.0061.

Burrows, R. (2012). Living with the h-index? Metric assemblages in the contemporary academy. *The Sociological Review*, 60(2):355–372. doi:10.1111/j.1467-954X.2012.02077.x.

Ceci, S. J. und Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8):3157–3162. doi:10.1073/pnas.1014871108.

Chase, J. M. (1970). Normative Criteria for Scientific Publication. *The American Sociologist*, 5(1):262–265.

Chavalarias, D. (2016). What's wrong with Science? *Scientometrics*, 110(1):1–23. doi:10.1007/s11192-016-2109-9.

Committee for the National Plan for the Future of the Humanities (2009). *Sustainable Humanities: Report from the National Committee on the Future of the Humanities in the Netherlands*. Amsterdam: Amsterdam University Press. doi:10.5117/9789089641427.

Coryn, C. L. S., Hattie, J. A., Scriven, M. und Hartmann, D. J. (2007). Models and Mechanisms for Evaluating Government-Funded Research: An International Comparison. *American Journal of Evaluation*, 28(4):437–457. doi:10.1177/1098214007308290.

CWTS (2012). *Merit, expertise and measurement*. Leiden: CWTS. Retrieved February 28, 2013, from http://www.cwts.nl/pdf/cwts_research_programme_2012-2015.pdf.

Daniel, H.-D., Mittag, S. und Bornmann, L. (2007). The potential and problems of peer evaluation in higher education and research. In A. Cavalli (Hrsg.), *Quality assessment for higher education in Europe*, S. 71–82. London: Portland Press.

Derrick, G. und Samuel, G. (2017). The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability. *Palgrave Communications*, 3:17040. doi:10.1057/palcomms.2017.40.

Donovan, C. (2007). The qualitative future of research evaluation. *Science and Public Policy*, 34(8):585–597. doi:10.3152/030234207X256538.

Edwards, M. A. und Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, 34(1):51–61. doi:10.1089/ees.2016.0223.

Enserink, M. (2015). Dutch sexism study comes under fire. *Science*, 364(6446). doi:10.1126/science.aad4618.

Frantz, T. T. (1968). Criteria for publishable manuscripts. *Personnel and Guidance Journal*, 47:384–386.

Galleron, I., Ochsner, M., Spaapen, J. und Williams, G. (2017). Valorizing SSH research: Towards a new approach to evaluate SSH research' value for society. *Fteval Journal for Research and Technology Policy Evaluation*, 44:35–41. doi:10.22163/fteval.2017.274.

Geuna, A. und Martin, B. R. (2003). University Research Evaluation and Funding: An International Comparison. *Minerva*, 41(4):277–304. doi:10.1023/B:MINE.0000005155.70870.bd.

Geuna, A. und Martin, B. R. (2001). University Research Evaluation and Funding: An International Comparison. SPRU Electronic Working Paper Series (Vol. 71).

van Gestel, R. und Lienhard, A. (Hrsg.) (2019). *Evaluating academic legal research in Europe. The advantage of lagging behind*. Cheltenham: Edward Elgar.

Glänzel, W., Thijs, B. und Debackere, K. (2016). Productivity, performance, efficiency, impact—What do we measure anyway? *Journal of Informetrics*, 10(2):658–660. doi:10.1016/j.joi.2016.04.008.

Gozlan, C. (2016). Les sciences humaines et sociales face aux standards d'évaluation de la qualité académique. *Sociologie*, 7(3):261–21. doi:10.3917/socio.073.0261.

Guetzkow, J., Lamont, M. und Mallard, G. (2004). What is Originality in the Humanities and the Social Sciences? *American Sociological Review*, 69(2):190–212. doi:10.1177/000312240406900203.

Guillory, J. (2005). Valuing the humanities, evaluating scholarship. *Profession (MLA)*, 11:28–38.

Gulbrandsen, M. (2000). Research Quality and Organisational Factors: An Investigation of the Relationship. Norwegian University of Science and Technology (NTNU).

Gumpenberger, C., Glänzel, W. und Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, 108(2):977–982. doi:10.1007/s11192-016-1991-5.

Hamann, J. (2016). The visible hand of research performance assessment. *Higher Education*, 72(6):1–19. doi:10.1007/s10734-015-9974-7.

Hammarfelt, B. und de Rijcke, S. (2014). Accountability in context: effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. *Research Evaluation*, 24(1):63–77. doi:10.1093/reseval/rvu029.

Hartmann, I. und Neidhardt, F. (1990). Peer review at the Deutsche Forschungsgemeinschaft. *Scientometrics*, 19(5–6):419–425.

Hemlin, S. (1993). Scientific quality in the eyes of the scientist. A questionnaire study. *Scientometrics*, 27(1):3–18.

Hemlin, S. und Montgomery, H. (1990). Scientists conceptions of scientific quality: An interview study. *Science Studies*, 1:73–81.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2):251–261. doi:10.1016/j.respol.2011.09.007.

Hicks, D. (2010). Overview of models of performance-based research funding systems. In *Performance-based Funding for Public Research in Tertiary Education Institutions*, S. 23–52. OECD Publishing. doi:10.1787/9789264094611-4-en.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glanzel und U. Schmoch (Hrsg.), *Handbook of quantitative science and technology research*, S. 473–496. Dordrecht: Kluwer Academic.

Hornung, A., Khlavna, V. und Korte, B. (2016). Research Rating Anglistik/Amerikanistik of the German Council of Science and Humanities. In M. Ochsner, S. E. Hug und H.-D. Daniel (Hrsg.), *Research Assessment in the Humanities. Towards Criteria and Procedures*, S. 219–233. Cham: Springer International Publishing. doi:10.1007/978-3-319-29016-4_17.

Hose, M. (2009). Qualitätsmessung: Glanz und Elend der Zahl. Retrieved January 13, 2012, from http://hsozkult.geschichte.hu-berlin.de/forum/id=1115&type=diskussionen.

Hug, S. E. und Aeschbach, M. (2020). Criteria for assessing grant applications: a systematic review. *Palgrave Communications*, 6(37). doi:10.1057/s41599-020-0412-9.

Hug, S. E., Ochsner, M. und Daniel, H.-D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal of Education Law and Policy*, 10(1):55–68.

Hug, S. E., Ochsner, M. und Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: a Delphi study among scholars of English literature, German literature and art history. *Research Evaluation*, 22(5):369–383. doi:10.1093/reseval/rvt008.

Johnston, R. (2008). On structuring subjective judgements: originality, significance and rigour in RAE2008. *Higher Education Quarterly*, 62(1-2):120–147. doi:10.1111/j.1468-2273.2008.00378.x.

Jonkers, K. und Zacharewicz, T. (2016). *Research Performance Based Funding Systems: a Comparative Assessment (No. EUR 27837 EN)*. Brussels: European Commission.

Kekäle, J. (2002). Conceptions of quality in four different disciplines. *Tertiary Education and Management*, 8(1):65–80. doi:10.1023/A:1017947005085.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Oxford: World Book Co.

Kerr, S., Tolliver, J. und Petree, D. (1977). Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1):132–141.

KNAW (2005). *Judging research on its merits. An advisory report by the Council for the Humanities and the Social Sciences Council*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

KNAW (2009). *Standard Evaluation Protocol 2009–2015: Protocol for Research Assessment in the Netherlands*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

KNAW (2011). *Quality Indicators for Research in the Humanities*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Lack, E. (2008). Einleitung. Das Zauberwort „Standards“. In E. Lack und C. Markschies (Hrsg.), *What the hell is quality? Qualitätsstandards in den Geisteswissenschaften*, S. 9–34. Frankfurt a. M.: Campus.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge: Harvard University Press.

Lanamäki, A., Ahmad, M. U. und Ochsner, M. (2019). Any publicity good publicity? The effect of satirical bias on Twitter and the Altmetrics Attention Score. In *Proceedings of the 3rd Conference on Research Evaluation in the Social Sciences and Humanities (RESSH), València*. València: Universitat Politècnica de València.

Langfeldt, L., Bloch, C. W. und Sivertsen, G. (2015). Options and limitations in measuring the impact of research grants—evidence from Denmark and Norway. *Research Evaluation*, 24(3):256–270. doi:10.1093/reseval/rvv012.

Langfeldt, L. (2010). Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation*, 13(1):51–62.

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation*, 15(1):31–41. doi:10.3152/147154406781776039.

Lawrence, P. A. (2002). Rank injustice. *Nature*, 415(6874):835–836. doi:10.1038/415835a.

van Leeuwen, T. N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie – Praxis und Forschung*, 2(8).

LERU (2012). Research Universities and Research Assessment. League of European Research Universities.

Leung, K. und Zhang, J. (1995). Systemic Considerations: Factors Facilitating and Impeding the Development of Psychology in Developing Countries. *International Journal of Psychology*, 30(6):693–706. doi:10.1080/00207599508246595.

Lienhard, A., Tanquerel, T., Flueckiger, A., Amschwand, F., Byland, K. S. und Herrmann, E. (2016). *Forschungsevaluation in der Rechtswissenschaft: Grundlagen und empirische Analyse in der Schweiz*. Bern: Stämpfli.

Lindsey, D. und Lindsey, T. (1978). The outlook of journal editors and referees on the normative criteria of scientific craftsmanship. *Quality and Quantity*, 12:45–62.

MacRoberts, M. H. und MacRoberts, B. R. (2017). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 10:646–649. doi:10.1002/asi.23970.

Mali, F., Pustovrh, T., Platinovšek, R., Kronegger, L. und Ferligoj, A. (2017). The effects of funding and co-authorship on research performance in a small scientific community. *Science and Public Policy*, 44(4):486–496. doi:10.1093/scipol/scw076.

Martens, S. und Schatz, W. (2016). Quality criteria and indicators for research in Theology: What to do with quantitative measures? In *21st International Conference on Science and Technology Indicators – STI 2016. Book of Proceedings*. València: Universitat Politècnica de València.

Merton, R. K. (1957). *Social Theory and Social Structure*. Glencoe, IL: Free Press.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer. doi:10.1007/1-4020-3714-7.

Moed, H. F. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity? *Scientometrics*, 74(1):153–161. doi:10.1007/s11192-008-0108-1.

Molinié, A. und Bodenhausen, G. (2010). Bibliometrics as Weapons of Mass Citation. *CHIMIA International Journal for Chemistry*, 64(1):78–89. doi:10.2533/chimia.2010.78.

Montada, L., Krampen, G. und Burkard, P. (1999). Personal and social orientations of psychology college teachers on evaluative criteria for own job performances: Results of an expert survey in German psychology college teachers. *Psychologische Rundschau*, 50(2):69–89.

Mutz, R., Bornmann, L. und Daniel, H.-D. (2014). Testing for the fairness and predictive validity of research funding decisions: A multilevel multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian science fund. *Journal of the Association for Information Science and Technology*, 66(11):2321–2339. doi:10.1002/asi.23315.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1):81–100. doi:10.1007/s11192-006-0007-2.

Ochsner, M. und Dokmanović, M. (2017). Quality criteria and research obstacles in the SSH in Macedonia. In *Second international conference on research evaluation in the social sciences and humanities*, S. 69–71. Antwerp: RESSH.

Ochsner, M., Hug, S. E. und Daniel, H.-D. (2012a). Indicators for Research Quality in the Humanities: Opportunities and Limitations. *Bibliometrie – Praxis und Forschung*, 1(4).

Ochsner, M., Hug, S. E. und Daniel, H.-D. (2012b). Wie wollen und sollen die Geisteswissenschaften Qualität und Leistung messen und steuern? In SAGW (Hrsg.), *Für eine neue Kultur der Geisteswissenschaften?*, S. 157–171. Bern: SAGW.

Ochsner, M., Hug, S. E. und Daniel, H.-D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation*, 22(2):79–92. doi:10.1093/reseval/rvs039.

Ochsner, M., Hug, S. E. und Daniel, H.-D. (2014). Setting the stage for the assessment of research quality in the humanities. *Consolidating the results of four empirical studies. Zeitschrift für Erziehungswissenschaft*, 17(6):111–132. doi:10.1007/s11618-014-0576-4.

Ochsner, M., Hug, S. E. und Daniel, H.-D. (2017). Assessment Criteria for Early Career Researcher's Proposals in the Humanities. In *21st International Conference on Science and Technology Indicators – STI 2016. Book of Proceedings*, S. 105–111. València: Universitat Politècnica de València.

Ochsner, M., Hug, S. E. und Galleron, I. (2017). The future of research assessment in the humanities: bottom-up assessment procedures. *Palgrave Communications*, 3:17020. doi:10.1057/palcomms.2017.20.

Ochsner, M., Kulczycki, E. und Gedutis, A. (2018). The Diversity of European Research Evaluation Systems. In P. Wouters, R. Costas, T. Franssen und A. Yegros-Yegros (Hrsg.), *Proceedings of the 23rd International Conference on Science and Technology Indicators, Leiden*, S. 1234–1241. Leiden: Leiden University.

Plag, I. (2016). Research Assessment in a Philological Discipline: Criteria and Rater Reliability. In M. Ochsner, S. E. Hug und H.-D. Daniel (Hrsg.), *Research Assessment in the Humanities. Towards Criteria and Procedures*, S. 235–247. Cham: Springer International Publishing. doi:10.1007/978-3-319-29016-4_18.

Plumpe, W. (2009). Qualitätsmessung: Stellungnahme zum Rating des Wissenschaftsrates aus Sicht des Historikerverbandes. Retrieved January 13, 2012, from http://hsozkult.geschichte.hu-berlin.de/forum/id=1101&type=diskussionen.

Plumpe, W. (2010). Der Teufel der Unvergleichbarkeit. Über das quantitative Messen und Bewerten von Forschung. *Forschung und Lehre*, 17(8):572–574.

Polanyi, M. (1962). The Republic of Science: Its Political and Economic Theory. *Minerva*, 1(1):54–73.

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P. und Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, 25(2):161–169. doi:10.1093/reseval/rvv038.

Riordan, P., Ganser, C. und Wolbring, T. (2011). Measuring the quality of research. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 63(1):147–172. doi:10.1007/s11577-010-0126-x.

Rolfe, G. (2013). *The university in dissent. Scholarship in the corporate university*. Abingdon: Routledge.

Rowney, J. A. und Zenisek, T. J. (1980). Manuscript characteristics influencing reviewers' decisions. *Canadian Psychology*, 21(1):17–21.

Schiene, C. und Schimank, U. (2006). Forschungsevaluation als Organisationsentwicklung: die Wissenschaftliche Kommission Niedersachsen. *Die Hochschule*, 15(1):46–62.

Schmidt, U. (2005). Zwischen Messen und Verstehen. Anmerkungen zum Theoriedefizit in der deutschen Hochschulevaluation. *evaNet-Positionen*, 6/2005.

Schmidt, U. (2010). Wie wird Qualität definiert? In M. Winde (Hrsg.), *Von der Qualitätsmessung zum Qualitätsmanagement: Praxisbeispiele an Hochschulen*, S. 10–17. Essen: Edition Stifterverband, Verwaltungsgesellschaft für Wissenschaftspflege. Retrieved from http://stifterverband.info/veranstaltungen/archiv/2010/2010_11_26_qualitaetsmanagement_an_hochschulen/index.html.

Schmied, M., Byland, K. und Lienhard, A. (2018). Procedures and criteria for evaluating academic legal publications: Results of a survey in Switzerland. *Research Evaluation*, 27(4):335–346. doi:10.1093/reseval/rvy020.

Sousa, S. B. und Brennan, J. L. (2014). The UK Research Excellence Framework and the Transformation of Research Production. In C. Musselin und P. N. Teixeira (Hrsg.), *Reforming Higher Education: Public Policy Design and Implementation*, S. 65–80. Dordrecht: Springer. Dordrecht. doi:10.1007/978-94-007-7028-7_4.

Sternberg, R. J. (2003). *The Psychologist's Companion. A guide to scientific writing for students and researchers*. Cambridge, UK: Cambridge University Press, 4. Aufl.

Sternberg, R. J. und Gordeeva, T. (1996). The Anatomy of Impact: What Makes an Article Influential? *Psychological Science*, 7(2):69–75. doi:10.1111/j.1467-9280.1996.tb00332.x.

Tahamtan, I. und Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12(1):203–216. doi:10.1016/j.joi.2018.01.002.

Tamblyn, R., Girard, N., Qian, C. J. und Hanley, J. (2018). Assessment of potential bias in research grant peer review in Canada. *Canadian Medical Association Journal*, 190(16):E489–E499. doi:10.1503/cmaj.170901.

Thorngate, W., Dawes, R. M. und Foddy, M. (2009). *Judging merit*. New York, NY: Psychology Press.

von Tunzelmann, N. und Kraemer Mbula, E. (2003). Changes in research assessment practices in other countries since 1999: final report.

Tversky, A. und Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131. doi:10.1126/science.185.4157.1124.

VSNU, NWO, KNAW (2003). *Standard Evaluation Protocol 2003–2009 for Public Research Organisations*. Utrecht: VSNU, NWO, KNAW.

VSNU, NWO, KNAW (2016). *Standard Evaluation Protocol 2015–2021. Protocol for Research Assessments in the Netherlands*. Voorburg: VSNU, NWO and KNAW.

Wissenschaftsrat (2004). Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung. Hamburg.

Wissenschaftsrat (2008a). Pilotstudie Forschungsrating Chemie. Köln.

Wissenschaftsrat (2008b). Pilotstudie Forschungsrating Soziologie. Köln.

Wissenschaftsrat (2010). *Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften*. Köln: Wissenschaftsrat.

Wissenschaftsrat (2011). Bewertungsmatrix. Forschungsrating Anglistik/Amerikanistik (Stand: September 2011). Retrieved January 19, 2012, from http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating/dokumente/.

Wolff, W. M. (1970). A study of criteria for journal manuscripts. *American Psychologist*, 25(7):636–639. doi:10.1037/h0029770.

Zacharewicz, T., Lepori, B., Reale, E. und Jonkers, K. (2018). Performance-based research funding in EU Member States—a comparative assessment. *Science and Public Policy*, 46(1):105–115. doi:10.1093/scipol/scy041.