Über Nachteile von Vorteilen

Ein Kommentar zu Richard Münch: „Die Konstruktion soziologischer Exzellenz durch Forschungsrating", in: Soziale Welt 60, S. 63-89

Von Friedhelm Neidhardt

Die Bewertungsgruppe Soziologie, die für das Forschungsrating des Wissenschaftsrats verantwortlich war, drückt ihr Verständnis vom Stellenwert der unternommenen Pilotstudie im Abschlussbericht des Wissenschaftsrats mit folgender Bemerkung aus: „Die Offenlegung nicht nur der ermittelten Befunde, sondern auch der dabei eingesetzten Methoden und der angefallenen Zwischenergebnisse dienen dem Zweck, das Forschungsrating für speziell Betroffene und allgemein Interessierte nachvollziehbar und kritisch diskutierbar zu machen. Die Pilotstudie sollte nur als erster Schritt einer Entwicklung verstanden werden, die die Bewertungsgruppe als unumkehrbar ansieht. „‚Rating' wird als Teil evaluativer Prozesse im Wissenschaftssystem nicht verschwinden. Die nachhaltige Qualität des Verfahrens und die Förderlichkeit ihrer Effekte werden davon abhängen, dass sie Teil eines Lernprozesses sind, an dem auch die Bewerteten mit ihren Fachgemeinschaften selbst kritisch mitwirken." (Wissenschaftsrat 2008: 346)

Richard Münch, selber Mitglied der Bewertungsgruppe, hat dieses Postulat ernst genommen, und er legt zur „Konstruktion soziologischer Exzellenz durch Forschungsrating" eine Analyse vor, die die bisherigen Ansätze kritisch aufnimmt und mit aufwendigem Zahlenwerk interessant weiterführt. Dazu gibt es viele Ansatzpunkte. Denn in der Tat ergaben sich die „Erhebungspraktiken, Datenoperationalisierungen, Indikatorenkalküle und Skalendefinitionen" aus „Güteabwägungen, die im Einzelnen auch anders hätten ausfallen können, als sie unter den wahrgenommenen Umständen im Forschungsrating balanciert wurden" (Neidhardt 2008: 429 f). Durch Einlassungen und Anregungen wie den von Münch vorgelegten kann das Verfahren, wenn man es wiederholen will, nur gewinnen – auch dann, wenn man nicht alle Einlassungen richtig findet und nicht allen Anregungen folgen will.

1. Einwände zum Forschungsrating

Schon der auf fünf Jahre fixierte *Untersuchungszeitraum* des Forschungsratings ist eine vom Wissenschaftsrat beschlossene Festlegung, für die man sich gut andere Grenzziehungen vorstellen kann; die Bewertungsgruppe neigte zu der Empfehlung, den Zeitraum zu verlängern – und ich selber bin dezidiert dafür. Aber auch dann könnten manche Misslichkeiten nicht verhindert werden: Ein wichtiger Artikel ist zum Endzeitpunkt zwar geschrieben, aber noch nicht veröffentlicht worden; ein Drittmittelprojekt ist schon bewilligt, doch noch nicht gestartet; ein hochgeschätzter und erfolgreicher Kollege ist kurz vor dem Stichtag pensioniert worden. Gerechte Vergleiche setzen immer die Markierung gleicher Bezugspunkte voraus, und wenn es dabei um Zeitbestimmungen geht, kommt man um die Definition von Stichtagen nicht herum. Wer das zu lange kritisiert, gerät schnell ins Lamentieren.

Gravierend sind die Probleme, die sich bei der Erfassung der *Erhebungs- und Analyseeinheiten* des Forschungsratings vor allem bei der Bestimmung von „Forschungsqualität" ergeben (Wissenschaftsrat 2008: 376 f). Sie entstehen als Folge des Vorsatzes, nicht nur das zu tun, was bei gängigen Ratings und Rankings sonst durchweg geschieht, nämlich die Leistung eines Faches an seinem Ort pauschal zu bestimmen – als stellten die einzelnen Disziplinen an den Universitäten in sich eine Forschungseinheit „mit jeweils gemeinsamer Forschungsplanung und abgestimmter Projektgestaltung" dar. Es gehört zu den Ergebnissen der Wissenschafts-

Soziale Welt 60 (2009), S. 325 – 332

ratsstudie, dass das Leistungsniveau der einzelnen Soziologien an ihrem jeweiligen Ort sehr stark variiert (Wissenschaftsrat 2008: 442 f). Will man deshalb aus guten Gründen stärker disaggregieren, entstehen aber erhebliche Folgeprobleme.

Ungereimtheiten ergaben sich nicht zuletzt dadurch, dass den Forschungseinrichtungen die Bestimmung ihrer Forschungseinheiten nicht von außen vorgegeben werden sollte, ihnen also Spielräume blieben, die sie in einigen Fällen in der Absicht nutzten, ihre Schwachstellen zu verstecken. Die Effekte machen sich in der Analyse spätestens dann bemerkbar, wenn für die Veröffentlichung der Befunde gewogene Mittelwerte berechnet werden, um die aus datenschutzrechtlichen Gründen nicht nennbaren einzelnen Forschungseinheiten wieder zusammenzufassen – ein erheblicher Informationsverlust, der auch, wie Münch (2009: 64 f) zeigt, mit der Möglichkeit von Verzerrungen verbunden ist.[1] Zusätzliche Differenzierungen sorgen nachfolgend für Aggregationsprobleme, die bei allen sonstigen Ratings und Rankings nicht entstehen, weil sie ihren Untersuchungsgegenstand von vornherein nicht differenzieren. Dass der von Münch beklagte Umstand, dass die Streuung individueller Leistungsqualitäten von kollektiven Gesamtwerten überdeckt wird, hier noch stärker auftritt als in den Darstellungen des Wissenschaftsrats (2008: 452ff), sollte allerdings nicht unterschlagen werden.

Unabhängig von diesen Aggregationsproblemen stellen sich Schwierigkeiten dar, die sich auf der Ebene der Forschungseinheiten bei der Messung von Forschungsqualität dadurch ergaben, dass nicht für alle relevanten Indikatoren valide Daten vorlagen. Dies führt in einer reinen Kennziffernarithmetik zu einer problematischen Übergewichtung des Indikators „*Peer Review Artikel in Fachjournalen*" – ein Sachverhalt, dessen Effekt Richard Münch allerdings deshalb stark überschätzt, weil er die auch mit seiner Hilfe angestrengten Abwägungen der Bewertungsgruppe, in die zahlreiche Indikatoren eingehen, zwar allgemein erwähnt, aber nicht dort berücksichtigt, wo sie wichtig waren (vgl. Wissenschaftsrat 2008: 364 f; Neidhardt 2008: 427ff). Für die Behauptung, „um in der FQ hohe Punktwerte zu erzielen", sei „nahezu ausschließlich das folgende Leistungskriterium entscheidend: die Zahl der PR-Artikel, relativ zum Personal aus Grund- und Drittmittelausstattung" (Münch 2009: 68), verkennt das breite Spektrum der Gesichtspunkte, welche die Bewertungsgruppe zu berücksichtigen hatte – und auch, je nach Fallbesonderheiten mal mehr, mal weniger, berücksichtigt hat. Die „Bewertungsmatrix" des Forschungsratings (Wissenschaftsrat 2008: 405 f) zeigt das Vielerlei an, das für die gemeinsam verfügten Endnoten zu kalkulieren und zusammenzufassen war: die Beurteilung eingereichter Publikationen, der Listen sowohl von Publikationen als auch von eingeworbenen Drittmittelprojekten, auch der „Selbstdarstellung der eigenen Stärken und Schwächen" – und quantitativ: neben der Zahl der Artikel in Fachzeitschriften mit Peer Review und einigen Hintergrundinformationen auch die „Zahl wissenschaftsgesteuert zugewiesener Drittmittelprojekte", dies alles in absoluten und relativen Ausprägungen. Gewiss hat Richard Münch recht mit der Feststellung, dass bei der Einschätzung von Forschungsqualität nicht alle Indikatoren bei den Gutachtern den gleichen Rang und die „PR-Artikel" ein herausragendes Gewicht besaßen. Nicht richtig ist aber seine Feststellung, PR-Artikel hätten „etwa zwei Drittel der Varianz in den Qualitätsurteilen des Forschungsratings erklärt" (Münch 2009: 70); wahrscheinlich ist fälschlicherweise ein einfacher Korrelationskoeffizient in diesem Sinne interpretiert worden. Aus multiplen Regressionsanalysen der Geschäftsstelle des Wissenschaftsrats ergibt sich etwas anderes. In einem Gesamtmodell mit den fünf wichtigsten Variablen (Determinations-Koeffizient 0,83) – ich habe mir das von der Geschäftsstelle des Wissenschaftsrats noch einmal bestätigen lassen – erklären die PR-Artikel nur rund ein Viertel der Bewertung, während

1 Das in der zweiten Pilotstudienrunde gegenwärtig laufende Forschungsrating von Elektrotechnik und Informationstechnik arbeitet aufgrund dieser Erfahrungen mit der Vorgabe eines Schemas von vier Forschungseinheiten pro Einrichtung, in die sich alle Forschenden nach eigener Wahl einordnen können.

die qualitativen Einschätzungen der Publikationen, die von den Forschungseinheiten eingereicht worden waren, sowie der ermittelten Publikations- und Drittmittellisten ein Spitzengewicht einnehmen, das gut die Hälfte des Ergebnisses erklärt.[2]

Diese Befunde relativieren die Bedeutung von PR-Artikeln gegenüber den Annahmen Münchs, sie widersprechen aber nicht grundsätzlich seinen Problematisierungen. Unstrittig ist, dass im Forschungsrating „nur eine kleine Minderheit der Publikationen bei den *quantitativen* Analysen durchweg und übergewichtet veranschlagt werden konnte" (Neidhardt 2008: 427) – und das ist in der Tat nicht befriedigend.[3]

2. Zahlenspiele

Besondere Verdienste erwirbt sich Richard Münch dadurch, dass er nicht nur kritisiert, sondern sich selber um die Verbesserung der im Forschungsrating eingesetzten Messinstrumente bemüht. Es geht ihm dabei insbesondere um den von der Bewertungsgruppe ausgedrückten Wunsch, „dass bei fortgesetztem Einsatz von Forschungsratings sowohl die Datennachfrage als auch die Indikatorenbildung noch stärker standardisiert wird" (Wissenschaftsrat 2008: 394). Bewundernswert akribisch sorgt er für eindrucksvolle Datenschöpfungen in großer Zahl. Nicht alle Operationen sind mir nachvollziehbar gewesen, und nicht bei allen Tabellenwerten hatte ich den Eindruck, sie seien korrekt. Darauf kommt es in diesem Zusammenhang aber weniger an als auf den Nachweis, dass in einem Forschungsrating zusätzliche Parametrisierungen prinzipiell möglich sind. Ich nehme im Folgenden zwei Beispiele auf, die ich stimulierend finde; sie betreffen den Umgang mit den Indikatoren „Reputation" und „Peer Review-Artikel".

Die Bewertungsgruppe Soziologie hatte in den „Empfehlungen zu den einzelnen Kriterien" am Schluss der vorgelegten Analyse zum Thema „*Reputation*" Folgendes vermerkt: „Da Reputation als ein sehr instruktiver Indikator gelten kann, sollten die auf ihn bezogenen Fragestellungen bei künftigen Erhebungen spezifischer dimensioniert und eindeutiger definiert werden. Dann scheint es auch möglich, im Hinblick auf eine Reihe von Daten (Herausgeber-, Beirats-, Fachgutachtertätigkeiten, ebenso bestimmte Arten von Ehrungen) Standardisierungen und Quantifizierungen vorzunehmen. Deren Aggregation setzt allerdings Gewichtungen voraus, über deren Festlegungen sich lange streiten lässt." (Wissenschaftsrat 2008: 400) Richard Münch nimmt diese Anregung (allerdings ohne Berücksichtigung von „Ehrungen")[4] auf, und er scheut sich auch nicht, die im Forschungsrating erfassten Ämter mit genauen Punkt-

2 Dass den PR-Artikelwerten nicht die übergroße Bedeutung zukam, die Richard Münch für sie behauptet, lässt sich am unteren Ende der Notenskala auch an folgendem Befund erkennen: Da 61 Forschungseinheiten im Untersuchungszeitraum keinen einzigen PR-Artikel ausweisen konnten, hätten 61 Forschungseinheiten die Note 1 („nicht befriedigend") bekommen müssen, wäre dieses Kriterium voll durchgeschlagen; es waren aber nur 22 (also statt 24 % nur 9 %), die so schlecht abschnitten.

3 Das Hauptproblem ergibt sich in diesem Zusammenhang aus der Unterschätzung von Monographien (vgl. Clemens et al. 1995). Sie resultiert daraus, dass die Verlage ihren Veröffentlichungsentscheidungen (anders als die renommierten Fachzeitschriften) nicht durchweg eine fachliche Qualitätskontrolle vorschalten. Buchpublikationen sind deshalb nicht schon per se als professionelle Leistungsnachweise interpretierbar. Eine Lösung des Problems dürfte sicher nicht darin bestehen, Monographien nach der Zahl ihrer Seiten zu gewichten und entsprechend differenziert in die Rechnungen einzubringen – sowie das CHE-Ranking geschieht. Wie die Monographiedaten in Tabelle 1 von Münch berechnet wurden, wird nicht mitgeteilt (Münch 2009: 66ff).

4 „Ehrungen", zum Beispiel mit Preisen, Ehrendoktoraten, Akademiemitgliedschaften etc., nehmen inzwischen auch in der Soziologie so zu, dass es sich lohnen könnte, sie in eine Reputationsrechnung mit aufzunehmen. Im Forschungsrating wurden für den Erhebungszeitraum insgesamt 194 einschlägige Angaben gemacht. 48 der 57 erfassten Forschungseinrichtungen nannten mindestens eine Auszeichnung eines ihrer Mitglieder (die Angaben von zwei Einrichtungen waren in dieser Hinsicht sogar zweistellig).

werten ungleich zu bewerten und in Punkt- und Skalenwerte umzurechnen (Münch 2009: 71 f). Darüber lässt sich nun in der Tat im Einzelnen streiten, aber ich finde seine Operationalisierungen im Prinzip vertretbar. So etwa könnte man es (um Ehrungen ergänzt) machen, wenn man die Zuschreibung von Reputation in Leistungsbestimmungen quantitativ aufwerten will.[5]

Durchaus erwägenswert erscheinen mir auch die Qualifizierungen, die Richard Münch im Hinblick auf die wichtige Variable „*Peer Reviewed Articles*", wieder mit fast verwirrendem Aufwand, versucht (Münch 2009: 73ff). Völlig berechtigt finde ich seine Bemühungen um eine Verbesserung der in dieser Hinsicht in der Tat nicht sehr zuverlässigen Datenbasis. In dieser Hinsicht aber erlahmt sein Elan auf der begonnenen Korrekturstrecke, sodass in spätere Rechnungen überwiegend die von Informationszentrum Sozialwissenschaften gelieferten Daten eingehen. Die zusätzlichen Differenzierungen nach dem jeweils mehr oder weniger professionellen Status der Zeitschriften, in denen PR-Artikel nachgewiesen waren, sind für weitere Diskussionen sicher interessant, aber es ist fraglich, ob sich dabei auch die unterschiedlichen Gewichtungen, die Münch auf sie bezieht, als haltbar erweisen würden. Münch kommt mit seinen Perfektionierungsversuchen nicht aus den tückischen Gewinn- und Verlustrechnungen heraus, die sich bei fortgesetzten Komplexitätssteigerungen immer einstellen: Jeder Schritt wirft neue Fragen und Probleme auf.

Das spricht nun allerdings nicht unbedingt gegen die Möglichkeit, auf diese Weise doch Fortschritte zu machen. Die Frage aber ist, ob diese mit den Münchschen Operationen am Ende tatsächlich erreicht werden. Mit den vorliegenden Daten lassen sich darauf nur ungefähre Antworten geben, und ich will da sehr vorsichtig sein. Es lässt sich mit ihnen einerseits abschätzen, ob und wie viel sie gegenüber dem Forschungsrating des Wissenschaftsrats überhaupt verändern. Und danach lässt sich andererseits fragen, ob diese Veränderungen, erscheinen sie als signifikant, die Qualität besitzen, einen "Fortschritt" bedeuten.

Münch hat seine neuen Ansätze vor allem in die Berechnung des Kriteriums „*Impact*" (Münch 2009: 74ff) eingebracht, verstanden als „Beitrag [einer gesamten Forschungseinrichtung] zur Entwicklung der Wissenschaft im Fachgebiet und darüber hinaus" (Wissenschaftsrat 2008: 407 f). Eine ausladende Tabelle bringt neben Angaben über die im Forschungsrating des Wissenschaftsrats erreichten Impactwerte in 11 Spalten für 55 Forschungseinrichtungen der Soziologie eine Unmenge eigener Daten.[6] Bemerkenswert ist nun, dass Münch diese Werte nicht benutzt, um eigene Notenrechnungen vorzustellen, die mit denen des Wissenschaftsrats vergleichbar wären. Dies aber wäre die einzige Möglichkeit, Effekte festzustellen, die zu einer durchschlagenden Verbesserung nachfolgender Forschungsratings führen könnten. Es reicht für die Relevanz der Rechnungen von Münch nämlich nicht der von ihm beabsichtigte Nachweis aus, „wie durch Indikatorenwahl eine Rangfolge konstruiert wird" (Münch 2009: 68), denn der Wissenschaftsrat weist keine Rangfolgen aus: Das Forschungsrating ist „Rating", nicht „Ranking".

Der Unterschied zwischen beiden Messausweisen besteht darin, dass die Perzentilskala, mit der Zwischenrechnungen auch im Forschungs*rating* sehr differenziert angestellt wurden, auf

5 Mich wundert deshalb, dass Münch bei späteren Rechnungen zum Kriterienbereich „*Effizienz*" (Münch 2009: 80ff) seine Reputationswerte nicht einbezieht. Da dabei auch andere Indikatoren (siehe Wissenschaftsrat 2008: 408) ohne Begründung nicht berücksichtigt werden, erscheint mir der Informationswert der Tabelle 6 relativ gering.

6 Die Datenaufbereitung Richard Münchs ist übrigens insofern heikel, als sie gegen Auflagen verstößt, an die sich der Wissenschaftsrat selber halten musste, nämlich aus datenschutzrechtlichen Gründen genaue Zwischenmesswerte für die einzelnen Indikatoren eines Kriterienbereichs nicht auszuweisen und darüber hinaus einerseits für die Forschungs*einheiten* keine Noten und andererseits für die Forschungs*einrichtungen* nur Noten – und zwar auf- und abgerundete – zu veröffentlichen.

eine 5er-Notenskala reduziert wird, wenn es darum geht, vertretbare Aussagen zu veröffentlichen. Die Notenskala mindert im Vergleich zu der überprägnanten Perzentilskala die Genauigkeitsanforderungen ganz erheblich. Sie absorbiert die meisten Perzentildifferenzen, die man zur Bildung von „Rangfolgen" brauchen würde, aber nur scheingenau sind.

Nun ist es schwierig, ein Notenbildung von außen nachzuholen, die Münch selber wohl deshalb zu riskant war, weil sein auf Kennzifferenstatistik reduziertes *Ranking* nicht die qualitativen Merkmale einschließt, die im Forschungsrating von der Bewertungsgruppe mitbedacht wurden. Eine genauere Lektüre der Inhalte von Tabelle 5 gestattet aber doch zwei Beobachtungen. (1) Die vier „mittleren Skalenwerte", die Münch mit seinen Ansätzen alternativ berechnet hat, weisen fast keine Varianzen aus, die zu unterschiedliche Noten führen würden. (2) In der weit überwiegenden Zahl der Fälle würden sie auch zu keinen anderen Noten führen, als sie das Forschungsrating erbracht hat.

Immerhin ist aber die Zahl der gegenüber dem Forschungsrating wahrscheinlich abweichenden Fälle mit etwa 17 von 55 nicht gering, und auch wenn es sich in allen diesen Fällen (mit einer Ausnahme) nur um die Differenz von einer Note handelt und wenn man zusätzlich davon ausgeht, dass auch diese in den meisten Fällen verschwinden würde, wenn man die weiteren Tabellenwerte mit zusätzlich qualitativen Abwägungen berücksichtigen würde – selbst dann könnte eventuell ein Rest an differenten Fällen stehen bleiben, der deshalb ernst zunehmen wäre, weil die Fehlertoleranz bei Forschungsratings angesichts der möglichen Brisanz ihrer lokalen Effekte sehr gering sein sollte.

Betrachtet man mit Gregory Bateson (1983: 582) eine Information als „einen Unterschied, der einen Unterschied ausmacht", so ließe sich wohl sagen, das Ergebnis der Münchschen Tabelle 5 sei nicht sehr, aber doch etwas informativ.[7] Auf der einen Seite überrascht mit Blick auf die Notenergebnisse des Wissenschaftsrats das hohe Maß ihrer Robustheit. Auf der anderen Seite bleibt wohl ein Rest, bei dem sich zu fragen lohnt, ob die Unterschiede eine Überlegenheit der Münchschen Korrekturen indizieren oder nicht.

3. Kritische Rückfragen

Der Autor macht eine Prüfung der Überlegenheit seiner Operationen deshalb schwer, weil sein Text einerseits reich an Daten, andererseits arm an Begründungen für jene Operationalisierungen ist, denen sie sich verdanken. Validitätsfragen kommen immer wieder auf, ohne dass er sich ihrer annähme. Ich bringe ein Beispiel, bei dem mir das Vorgehen von Richard Münch überhaupt nicht einleuchtet.

In den wichtigen Tabellen 1 und 5, die zur Ermittlung von *Forschungsqualität* und *Impact* dienen, benutzt Richard Münch bei der Konstruktion zentraler Leistungswerte von Forschungseinrichtungen in mehreren Fällen Kennzahlen nur für die jeweils besten bzw. die drei besten Forschungseinheiten (Münch 2009: 66ff, 77ff).[8] Sein erklärtes Ziel ist zu zeigen, dass sich die Rangwerte der Forschungseinrichtungen je nach Messmethode verändern – und dass der grobe gewogene Gesamtmittelwert, den der Wissenschaftsrat anstelle der ermittelten Ein-

7 Das dürfte auch auf die Tabelle 8 zur *Nachwuchsförderung* zutreffen (Münch 2009: 83ff), obwohl der Autor in seinen Berechnungen die bewusst ungleiche Gewichtung der fünf Faktoren im Forschungsrating ohne Begründung nicht übernommen hat. Für die Differenzen zwischen dem eigenen „mittleren Skalenwert" und dem Ergebnis des Forschungsratings („Note FR") vermerkt er selber, dass seinen Berechnungen „die qualitative Beurteilung nach zusätzlichen Kriterien, die in die Notengebung des Forschungsratings eingegangen sind" fehlt (Münch 2009: 85). Auch deshalb sind die Zahlenwerte beider Spalten nicht vergleichbar.

8 Auf welche Weise für die Tabelle 5 die Forschungsqualität der drei besten Forschungseinheiten in den Fällen geschätzt wurden, in denen keine drei für eine Forschungseinrichtung angegeben worden waren, wird nicht mitgeteilt.

zelwerte zur Anonymisierung seines Datenberichts publiziert hat, in den Impactbestimmungen Streuungen verdeckt, die einen eigenen Informationswert besitzen. Die vom Autor nicht gestellte und dann auch nicht beantwortete Frage ist, welchen Sinn es unter Validitätsgesichtspunkten aber macht, anstelle des Gesamtmittelwerts für eine Forschungseinrichtung nur deren Spitzenreiter bzw. den Mittelwert der Spitzenreiter auszuweisen und dann diese Konstrukte in wiederum gemittelte Skalenwerte so einzurechnen, dass am Ende der Rechnung so etwas wie die Mitte der Spitze erscheint. Wozu soll das informativ sein? Und ist der schlechter bewertete Teil von Forschungseinheiten für die Einschätzung der Einrichtung eine *Quantité négligeable*?

Wie im Abschnitt 2 schon gezeigt, sind die Noteneffekte der verschieden Berechnungsarten Münchs zwar weit überwiegend gering; aber es gibt Sonderfälle, bei denen signifikante Abweichungen möglich oder gar wahrscheinlich sind – nämlich solche Fälle von Forschungseinrichtungen, bei denen die ausgewiesene Qualitätsstreuung der einzelnen Forschungseinheiten sehr groß ist.[9] Solche Fälle profitieren natürlich davon, dass sie nur auf der *sunny-side of the street* beleuchtet werden. Gerecht erscheint mir das nicht.

Um nicht selber allzu sehr ins analytische Klein-klein zu geraten, will ich im Folgenden nur noch einen grundsätzlichen Fall aufbringen, bei dem mir nicht die Feststellungen Münchs fraglich, aber deren Auslegungen schief erscheinen. Es geht um Kompromissbildungen im Forschungsprozess, zu denen jeder in der Sozialforschung gezwungen ist, wenn es darum geht, für bestimmte Operationen Vor- und Nachteile abzuwägen. Besondere Schwierigkeiten ergaben sich im Forschungsrating in dieser Hinsicht bei der Behandlung der beiden *Transferkriterien*, mit denen bestimmt werden sollte, wie leistungsstark soziologische Forschungseinrichtungen bei der Umsetzung von Erkenntnissen in gesellschaftliche Praxisfelder (Kriterium V: Beratungen, Dienstleistungen, sogen. Ausgründungen etc.) und bei deren Vermittlung mit Aufklärungs- und Bildungsangeboten (Kriterium VI: Weiterbildungsmaßnahmen, Beiträge in Zeitungen etc.) erscheinen.

Münch verweist bei der Behandlung dieser Komplexe zu Recht auf „unzureichende Daten" und „schwierige Quantifizierungen" (2009: 85ff). Er kommt an dieser Stelle mit eigenen Rechnungen auch nicht über das hinaus, was die Bewertungsgruppe im Forschungsrating gemacht hat.[10] Gleichwohl kritisiert er scharf die Folgerung, die die Bewertungsgruppe aus dem Mangel an handfesten Indizien gezogen hat, um den Genauigkeitsgrad ihrer Bewertung an den Informationsgehalt ihrer Daten anzupassen, nämlich bei beiden Transferkriterien die ansonsten fünfstellige Bewertungsskala auf eine dreistellige zu reduzieren. Dies bedeute, so Münch (2009: 86 bzw. 87), eine „Abwertung des Kriteriums" mit unerwünschten Effekten: „Sie bestraft diejenigen FER, die in diese Funktion investieren, und bevorteilt jene, die das unterlassen." Und das setze „ein Signal, das zum Abbau der entsprechenden Aktivitäten beiträgt."

Nun nehme ich an, dass die knapp 20 Forschungseinrichtungen, die bei den Transferkriterien öffentlich als „überdurchschnittlich" ausgezeichnet werden, dies nicht als eine Bestrafung empfinden. Dies umso weniger, als ihre Transferleistungen in anderen Ratings oder Rankings überhaupt nicht festgestellt und ausgewiesen werden. Eine Dreierdifferenzierung ist immerhin mehr als eine Nullaussage, und sie ist bekömmlich für den, der dabei oben landet. Insofern ist die Tatsache, dass der Wissenschaftsrat im Forschungsrating zum ersten Mal eigene Bewer-

9 Zu diesen Fällen gehört die Soziologie der Universität Bamberg. Sie stellt den einzigen Fall dar, bei dem für die Einschätzung der Forschungsqualität ihrer Forschungseinheiten die ganze Notenskala von 5 bis 1 gebraucht wurde (siehe Wissenschaftsrat 2008: 454).

10 Die Bewertungsgruppe macht in den Empfehlungen des Wissenschaftsrats Vorschläge, die bei künftigen Forschungsratings zu zusätzlichen Standardisierungen und zu einer Zusammenlegung beider Transferbereiche der Kriterien V und VI führen sollen – dies ausdrücklich mit dem Ziel, dann eine fünfstellige Skala einsetzen zu können (Wissenschaftsrat 2009: 403 f).

tungsdimensionen für Wissenschafts*anwendung* ausdifferenziert hat, gerade nicht „ein Signal, das zum Abbau entsprechender Aktivitäten beiträgt". Im Gegenteil: wer auf einer Dreierskala als „unterdurchschnittlich" ausgewiesen wird, erfährt – wenn überhaupt – eher den Druck, entsprechende Aktivitäten zu verstärken. Dieser Druck würde vielleicht noch größer sein, wenn auf einer Fünferskala das Prädikat „nicht befriedigend" verliehen würde. Aber, das bei der Kommentierung der Transferkriterien tatsächlich vorhandene „Manko des Forschungsratings" ist der Nachteil eines Vorteils. Es erscheint auf einem gesteigerten Analyseniveau. Die Kritik verliert Augenmaß, wenn sie die Vorteile hinter den Nachteilen übersieht.

4. „Lob der Inkonsequenz" (Leszek Kolakowski)

Richard Münch spürt natürlich auch die Nachteile der Vorteile, die er selber bei seinem *trial and error* erreicht (Münch 2009: 88). Sichtbar werden beim kritischen Durchgang durch seinen eigenen Text in der Tat einerseits Vorteile, die er durch zusätzliche Differenzierungen sowie durch couragierte Standardisierungen an einigen Stellen erreicht. Aber Forschungsratings führen nach seinen eigenen Worten in ein „vermintes Feld". Und Münch sieht bei sich selber das Problem, „die Konstruktion soziologischer Realität durch das Forschungsrating durch eine weitere Konstruktion zu verdoppeln" – nämlich durch die eigene Rechnerei. Und ihm ist offenkundig nicht wohl angesichts des „Dilemmas […], dass die hier durchgeführten Analysen immer noch das Spiel des Ratings mitmachen" – obwohl er diesem doch prinzipiell misstraut (unmissverständlich dazu Münch 2007).

Eigentlich – denkt man sich angesichts der Kautelen, mit denen Richard Münch sich seine eigenen Konstruktionen erträglich machen will – hätte es nahegelegen, dass er darauf verzichtet, ein Forschungsrating nicht nur zu kritisieren, sondern auch fortzuschreiben. Aber wir profitieren ja davon, dass er nicht konsequent genug ist, um den Gegenstand seiner prinzipiellen Abneigung nur anzugreifen. Seine Probegänge animieren, und sie zeigen in der Tat, dass wir etwa anderes herausbekommen, wenn wir anders messen – eine Trivialität, an die immer wieder zu erinnern lohnt.

Außerdem bleibt anzumerken, dass Richard Münch mit seiner Inkonsequenz ziemlich gut die Ambivalenzen ausdrückt, die fast alle Mitglieder der Bewertungsgruppe beim Forschungsrating empfanden. Man sah das Ungenügen der Kennziffernökonomie und hat doch nicht darauf verzichtet, sie ernst zu nehmen und für die eigenen Abwägungen zu benutzen. Im „Informed Peer Review" hat die Treffsicherheit des „Peer Review" auch sicher nicht darunter gelitten, mit Kennziffern „informed" zu sein. Gleichwohl ist bei den „Peers" angesichts der Unwägbarkeit der Folgen des eigenen Tuns nicht das Gefühl verschwunden, an einem heiklen Unternehmen beteiligt zu sein.

Und dennoch haben alle Mitglieder der Bewertungsgruppe zugestimmt, dass die eigenen Bewertungen vom Wissenschaftsrat ratifiziert und veröffentlicht – und dann auch wirksam werden. Maßgeblich dafür war die feste Annahme, dass in wechselnder Gestalt Evaluationen im Allgemeinen und Forschungsratings im Besonderen aus der Wissenschaft nicht verschwinden werden. Es kann für die betroffenen Disziplinen dann nicht gut sein, wenn ihre eigene Vermessung vorrangig den Interessen von Ministerien und Universitätsleitungen folgt oder den Marktabhängigkeiten privater Institute und Zeitschriften überlassen bleibt. „[…] academics should no longer leave evaluations to others, but should invest in self-defined measures of quality, relevance, and efficiency, and in the collection and propagation of data […]" (Schimank 2005: 375). Gefragt sind dann aber Kollegen und Kolleginnen, die in den fälligen Diskussionsprozessen der Profession so engagiert und einfallsreich Position beziehen, wie Richard Münch das tut.

Literatur

Bateson, G. (1983): Ökologie des Geistes, 2. Aufl., Frankfurt / Main.

Clemens, E. S. / W. W. Powell / K. McIlwaine / D. Okamoto (1995): Career in print. Books, journals, and scholarly reputations, in: American Journal of Sociology 101, S. 433-494.

Kolakowsky, L. (1967): Lob der Inkonsequenz, in: Der Mensch ohne Alternative. Von der Unmöglichkeit, Marxist zu sein, München, S. 214-223.

Münch, R. (2007): Die akademische Elite. Zur sozialen Konstruktion wissenschaftlicher Exzellenz, Frankfurt / Main.

Münch, R. (2009): Die Konstruktion soziologischer Exzellenz durch Forschungsrating, in: Soziale Welt 60, S. 63-89.

Neidhardt. F. (2008): Das Forschungsrating des Wissenschaftsrats. Einige Erfahrungen und Befunde, in: Soziologie 37 / 4, S. 421-432.

Schimank, U. (2005): ‚New public management' and the academic profession. Reflections on the German situation, in: Minerva 43, S. 361-367.

Wissenschaftsrat (2008): Dokumentation der Pilotstudie Soziologie. Teil III, in: Pilotstudie Forschungsrating. Empfehlungen und Dokumentation, Köln, S. 341-568.

Prof. em. Dr. Dr. h.c. Friedhelm Neidhardt
Wissenschaftszentrum Berlin
Reichpietschufer 50
10785 Berlin-Tiergarten
neidhardt@wzb.eu