Marc Torka

Institutioneller gleich handlungspraktischer Wandel? Das Beispiel von Begutachtungspraktiken bei der Evaluation wissenschaftlicher Einrichtungen

Verfahren der institutionellen Evaluation von ganzen wissenschaftlichen Einrichtungen und Forschungsfeldern erzeugen eine neue Handlungssituation für Gutachter/innen: Die Gegenstände, der Prozess und die soziale Situation der Urteilsfindung verändern sich im Vergleich zu anderen Formen der Begutachtung. Wandeln sich in dieser neuartigen Handlungssituation aber auch die Begutachtungspraktiken und Wertorientierungen von Gutachterinnen und Gutachtern? Oder werden diese nur auf neue Handlungssituationen übertragen? Um der Frage nachzugehen, inwieweit institutionelle Veränderungen der Wissenschaft auch zu neuen Handlungsweisen führen, werden Praktiken des Begutachtens im Kontext von zwei Evaluationstypen analysiert.

1 Das Problem, den institutionellen Wandel von Wissenschaft im Handeln nachzuweisen

Die institutionelle Umwelt der Wissenschaft unterliegt gegenwärtig einem umfangreichen Neuordnungsprozess. Vielfältige wissenschaftspolitische Initiativen zielen auf die Reorganisation von Studiengängen, Karrieren, Förderungsinstrumenten, Organisationsstrukturen und den hier diskutierten Bewertungssystemen der Wissenschaft ab. Hierfür sind die seit den 1980er Jahren vermehrt auftretenden Evaluationen von ganzen Wissenschaftsorganisationen und Wissenschaftsfeldern ein besonders intensiv diskutiertes Beispiel. Die vergangenen Leistungen und zukünftigen Potentiale von wissenschaftlichen Einrichtungen werden damit turnusmäßig überprüft, um über die Weiterfinanzierung und Fortsetzung von Forschungsprogrammen zu entscheiden. Für die Wissenschaftsforschung deuten solche von außen auferlegten Wissenschaftsevaluationen auf einen generellen Vertrauensverlust in die Selbststeuerungsmechanismen der Wissenschaft hin. Dieser mündet in einem gesteigerten Rechtfertigungsbedarf, dem mit Evaluationen begegnet wird (Weingart 2005, Schimank 2005). Bereits die Existenz von Wissenschaftsevaluationen liefert demnach einen Hinweis auf veränderte Autoritätsbeziehungen innerhalb der Wissenschaft und zwischen Wissenschaft und Politik (Whitley/Gläser/Engwall 2010).

Der vorherrschende wissenschaftssoziologische und wissenschaftspolitische Diskurs geht also von einer „New Balance of Power" aus und fragt vor allem danach, welche Folgen Evaluationen für die wissenschaftliche Praxis haben. Das Verhältnis zwischen institutioneller Gestalt und Praxis der Wissenschaft steht somit erneut zur Debatte.[1] Die dominante Vorstellung ist dabei, dass die Institutionalisierung von Evaluationsverfahren mehr oder minder ungebrochen zu einer Neuordnung wissenschaftlicher Wertorientierungen, Handlungs- und Bewertungsweisen führt. Evaluationen setzen formale Rahmenbedingungen und verwenden Entscheidungskriterien, an die sich das wissenschaftliche Handeln dann anpasst. So ziehen Ben Martin und Richard Whitley (2010) beispielsweise den Schluss, dass Wissenschaftsevaluationen den Wettbewerb um Publikationschancen und Forschungsmittel sowie die Herausbildung von disziplinären Eliten beförderten und deshalb generell mit einem „decline in collegiality" zu rechnen sei. Auch verkürzte Publikationsintervalle, die durch einen generellen Publikationsdruck, befristete Drittmittelforschung oder turnusmäßige Evaluationen erzeugt werden, brächten folglich „short term"-, „incremental"- und „mainstream"-

1 Das ist die klassische Konfliktlinie seit der Entstehung der Wissenschaftssoziologie. Die von Latour und Woolgar bzw. Knorr-Cetina begründeten mikrosoziologischen Laborstudien kritisierten seit den 1970er Jahren den zuvor dominanten institutionalistischen Ansatz von Merton, der die Norm- und Wertebasis, aber weniger das konkrete wissenschaftliche Handeln in den Blick nahm. Die „Erneuerung der institutionalistischen Wissenschaftssoziologie" (Schimank 1995) zielt hingegen eher auf rechtliche und formal organisatorische Regulierungen ab.

Forschungen hervor (ebd.: 70f.). Studien über die Grenzen des Einflusses von Evaluationen auf den Kern wissenschaftlichen Handelns sind hingegen rar.[2]

So plausibel und wichtig diese institutionalistische Forschungsperspektive ist, so gering ist bislang das empirisch gesicherte Wissen über die handlungspraktischen Folgen von Evaluationen und so groß sind die analytischen Herausforderungen zur Erforschung dieses Zusammenhangs (Gläser et al. 2008). Die Schwierigkeit besteht nämlich darin, „kausale Mechanismen" zu identifizieren und zu isolieren, die die Regeln von sporadischen Evaluationsereignissen in den Alltag wissenschaftlichen Handelns importieren (Gläser/Laudel 2007). Gelingt eine Isolierung von Mechanismen – zum Beispiel, Leistung nach Drittmittelquoten oder dem Impact von Publikationen zu berechnen, die Bereitstellung von Forschungsmöglichkeiten an diesen Kriterien auszurichten und auf diese Weise der Forschungspraxis einen Orientierungsrahmen für ‚werthaltige' Beiträge zu geben – dann bleibt die Frage offen, inwiefern diese Mechanismen ursächlich mit der Etablierung von Wissenschaftsevaluationen zusammenhängen.[3] Wie noch zu zeigen ist, sind Publikations- und Drittmittelraten noch nicht einmal die bestimmenden Kriterien bei der Urteilsfindung, sondern nur eine unter vielen Informationsquellen der Gutachter/innen.[4] Die institutionalistische Perspektive neigt außerdem dazu, die Reaktions- und Umgangsweisen von Wissenschaftler/innen mit neuartigen Evaluationsregimen unterzubelichten, obwohl diese „by no means just passive recipients of such changes in their institutional environment" (Leisyte/Enders/de Boer 2010: 267) sind. Wir werden sehen, dass gerade in der aktiven Auseinandersetzung mit neuen Regulierungsformen tief verankerte Eigenregulierungen von Wissenschaftler/innen zutage treten, die ihr Handeln strukturieren.

Vor dem Hintergrund dieser ungelösten Schwierigkeiten, die ‚Wirkung' gelegentlicher Evaluationsereignisse auf das Alltagshandeln von Wissenschaftler/innen zu analysieren, schlage ich im Folgenden einen bescheideneren Weg ein. Ich rücke eine Handlungssituation ins Zentrum, in der durch Evaluationen gestiftete Erwartungsstrukturen direkt auf Eigenregulierungen von Wissenschaftler/innen treffen: Anhand der Begutachtungsweisen von Wissenschaftlern im Rahmen von institutionellen Evaluationen frage ich, durch welche Regeln das Gutachterhandeln bestimmt ist. Dort ist es möglich, in situ und ohne verschiedene Handlungssituationen vermittelnde Mechanismen der Frage nachzugehen, inwieweit sich das Handeln von wissenschaftlichen Gutachtern dem institutionellen Rahmen, eigenen Kriterien der Angemessenheit oder einer Gemengelage aus beiden fügt. Da ich mit Praktiken der Urteilsfindung nur einen Aspekt wissenschaftlichen Handelns analysiere, können natürlich keine Schlussfolgerungen gezogen werden, ob sich das konkrete Forschungshandeln (z.B. die Problemwahl) an Evaluationskriterien ausrichtet oder nicht.

Anhand von zwei stark kontrastierenden Evaluationstypen gehe ich der Frage nach, ob sich im Kontext dieser neuen Handlungssituationen auch das Gutachterhandeln in substantieller Weise ändert und welcher Art diese Anpassung gegebenenfalls ist. Dafür werde ich zunächst auf die Evaluationstypen, die Datenbasis und die Methodik eingehen (2), dann der Frage nachgehen, inwieweit im Kontext dieser Evaluationen qua Verfahren (3) und entlang der Situationsdeutung von Gutachtern (4) überhaupt eine grundsätzlich neue Handlungssituation entsteht. Schließlich werde ich analysieren, mit welchen Begutachtungspraktiken diese gegebenenfalls neuen Herausforderungen bewältigt werden (5). Zum Schluss ziehe ich ein Fazit bezüglich der Frage, ob die Regeln der Evaluation sich im Handeln der Gutachter/innen niederschlagen und dieses modifizieren (6).

2 Für das Beispiel der Problemwahl vgl. jedoch Leisyte/Enders/de Boer 2010.

3 Stefan Hornbostel hat in seinem Tagungsbeitrag darauf aufmerksam gemacht, dass die Publikationsraten bereits im Zuge der Expansion des Wissenschaftssystems seit den 1960er Jahren rasant zugenommen haben und nicht erst mit der flächendeckenden Etablierung von Evaluationen ab Mitte der 1980er Jahre. Die Projektform als Grundlage der Drittmittelforschung ist ein noch älteres und in allen Wissenschaftsarten vorkommendes Phänomen (Torka 2009).

4 Jochen Gläser und Grit Laudel (2007) haben mit Australien einen besonders extremen Fall von Wissenschaftsevaluationen analysiert. Bis 2008 wurden ausschließlich metrische Daten verwendet. Seit 2010 findet sich aber auch dort ein dem britischen Research Assessment Exercise vergleichbares Peer Review-Verfahren.

2 Datenbasis, Evaluationstypen, Methodik

Bei meinen Ausführungen stütze ich mich auf reichhaltiges Datenmaterial, das im Kontext eines Forschungsprojekts über Urteilsprozesse in drei Evaluationsverfahren erhoben wurde.[5] Es handelt sich um das niederländische *Standard Evaluation Protocol*, das Evaluationsverfahren der deutschen Leibniz-Gemeinschaft und das britische *Research Assessment Exercise* von 2008. Mit allen diesen Verfahren sollen die Qualitäten wissenschaftlicher Organisationseinheiten überprüft und weiterentwickelt sowie Finanzierungsentscheidungen vorbereitet werden. Dennoch verfolgen sie die gleichen Ziele in unterschiedlicher Weise.

Es lassen sich zwei Grundtypen unterscheiden. Das britische Verfahren evaluiert alle nationalen wissenschaftlichen Einrichtungen zeitgleich, vergleichend und aus aktenkundiger Distanz und erzeugt letztlich ein numerisches Ranking, das zwar der Politik als Verteilungsschlüssel dient, aber kaum eine inhaltliche Rückmeldung an die Institute bietet. Dafür werden verschiedene Informationsquellen (v.a. einzelne „outputs" wie Publikationen oder Patente, aber auch Angaben zum „esteem" wie z.B. Preise oder die Wahl in bedeutsame Gremien und zum „environment" einer Institution, z.B. Stellen, Drittmittel, Dissertationen) an ein fachlich organisiertes Gutachterpanel gesendet. Aus der Einzelbenotung der heterogenen Informationen bildet dieses Team schließlich eine Gesamtnote, die den Leistungsvergleich zwischen wissenschaftlichen Einheiten (*Units of Assessment*) innerhalb eines Wissenschaftsgebiets ermöglicht.

Hingegen sehen die deutschen und niederländischen Verfahren Einzelfallbegutachtungen von wissenschaftlichen Einrichtungen vor. Entscheidungen werden nicht nur auf Grundlage von Akten, sondern in einem interaktiven Kontext erarbeitet (*Begehung* bzw. *site visit*). Am Ende spricht ein meist interdisziplinär zusammengesetztes Gutachterteam inhaltliche Empfehlungen zur Weiterentwicklung von Instituten aus. In Form einer Checkliste werden die Gutachter/innen dazu angehalten, heterogene Informationen über die Leistung und Leistungsfähigkeit einer je spezifischen wissenschaftlichen Einrichtung in ein Gesamtbild zu integrieren. Es werden jedoch keine Noten vergeben. Die niederländischen und deutschen Verfahren zielen deshalb nicht auf einen institutionellen Vergleich vergangener Leistungen ab, sondern sie tragen deutliche Züge einer Beratung hinsichtlich der zukünftigen Entwicklung einer gesamten wissenschaftlichen Organisation.

Für unsere Frage ist wichtig, dass die jeweiligen Verfahrensregeln zwar bestimmen, welche Informationen erbracht, beurteilt und am Ende kommuniziert werden sollen (numerisch/inhaltlich). Diese Regeln konstituieren die Typen einer vergleichend bewertenden und einer fallspezifisch beratenden Evaluation. Bei der Urteilsfindung selbst verfügen die Gutachter/innen jedoch über eine große Freiheit hinsichtlich der Auswahl, Gewichtung, Ausdeutung und Verwendungsweise von Gütekriterien. Die scientific community ist in den Evaluationsverfahren an zentraler Entscheidungsstelle positioniert und in ihrem Handeln nicht völlig durch Verfahrensregeln bestimmt.

Aus diesen Verfahren haben wir unterschiedliche Institute ausgewählt und deren Evaluationsprozess aus verschiedenen Perspektiven beleuchtet. Neben Dokumenten haben wir Interviews mit den jeweiligen Evaluationsagenturen, Institutsleitungen und -administrationen (vor und nach der Evaluation) und vor allem mit den beteiligten Gutachtern geführt und analysiert. Mit der prozessnahen und multiperspektivischen Erhebung sind wir dem methodischen Problem begegnet, dass uns das Kernereignis, die Kommunikation der Begutachtenden im Panel, verschlossen blieb.[6]

5 Es handelt sich um das Forschungsprojekt der Forschungsgruppe Wissenschaftspolitik am Wissenschaftszentrum Berlin für Sozialforschung: „Urteilsbildung im Peer Review. Internationale Fallstudien zur Evaluation von wissenschaftlichen Einrichtungen". Vgl. auch den Beitrag meiner Kolleginnen Silke Gülker und Dagmar Simon in diesem Band.

6 Mit dem Problem, die Gespräche in Gutachterpanels allenfalls beobachten, aber nicht aufzeichnen zu dürfen und damit diese Kommunikationsdynamik auch nicht rekonstruieren zu können, haben alle mir bekannten Studien zu

Im Zentrum der Interviews standen keine expliziten Deutungen, sondern hinreichend detaillierte Erzählungen über den Verlauf der konkreten Begutachtung, Erläuterungen der Vorgehensweise von Gutachterinnen und Gutachtern anhand konkreter Beispiele sowie Berichte über die dabei aufgetretenen Diskussionen und Probleme. Die Rekonstruktion zielte auf die Grundorientierungen, die Handlungen von Gutachtern zugrunde lagen.

3 Institutionelle Evaluation – Eine neue Handlungssituation?

Kommen wir zur ersten empirischen Teilfrage: Inwiefern entsteht im Kontext solcher institutionellen Evaluationen überhaupt eine neue Handlungssituation für Gutachter/innen? Das ist schon deshalb keine triviale Frage, weil Evaluationsverfahren die konkreten Einzelentscheidungen von Gutachter/innen nicht determinieren, Bewertungen von Forschungsleistungen zum Alltag von Wissenschaftler/innen gehören und bis in die Begriffsverwendung hinein Evaluationen mit dem wissenschaftseigenen Peer Review verschwimmen (z.B. Hirschauer 2002, Neidhard 2010). Welche Veränderungen lassen sich also ausmachen, wenn man die Verfahren institutioneller Evaluationen sowie die Situationsdeutungen von Evaluierten und Gutachtenden betrachtet? Das Verfahren lässt insbesondere drei Verschiebungen sichtbar werden.

In institutionellen Evaluationen sind nicht mehr einzelne Personen, Publikationen oder Forschungsvorhaben der *Begutachtungsgegenstand*, sondern die Gesamtleistung einer ganzen wissenschaftlichen Organisation tritt in den Vordergrund. Damit wandern auch zusätzliche Kriterien und Informationsquellen in die Evaluation ein, die dazu beitragen sollen, ein breites Gesamtbild der jeweiligen Einrichtung zu zeichnen und zu erkennen. Dazu gehören je nach Schwerpunkt des Verfahrens einzelne Publikationen, Listen über den Gesamtoutput, Personal-, Finanzierungs- und Drittmittelbilanzen, aber auch programmatische Selbstbeschreibungen der jeweiligen wissenschaftlichen Einrichtung. Auch wenn es keine verfahrensmäßigen Bestimmungen der Handhabe dieser heterogenen, teilweise neuartigen Informationsquellen gibt, verbinden sich hiermit verschiedene Verdachtsmomente. Das schwerwiegendste ist sicherlich, dass die Orientierung an quantitativen Indikatoren zunehme (z.B. Kieser 2010), um die Fülle an Einzelinformationen überhaupt bewältigen und zu einem Gesamtbild integrieren zu können. Da neben der wissenschaftlichen Qualität auch Fragen der geeigneten Infrastruktur sowie Leitungs- und Organisationsstruktur eine wichtige Rolle spielen, müssen Gutachter/innen zudem über Gegenstandsbereiche jenseits ihrer Fachexpertise urteilen. Es stellt sich deshalb die empirische Frage, welche Informationen für die Gutachter/innen in den jeweiligen Verfahren zentral sind, d.h. mit welchen Selektivitäten sie diesen neuen Gegenstand für sich handhabbar und beurteilbar machen (siehe 5.1).

Eine zweite Verschiebung betrifft den *Begutachtungsprozess*. Während bei der Begutachtung von Publikationen oder Forschungsvorhaben die Initiative stets bei den Begutachteten liegt, man etwas zu gewinnen hat und sich den jeweiligen Wettbewerb (Förderinstitutionen, Publikationsorte und -formate) aussuchen kann, ist das bei institutionellen Evaluationen nicht mehr der Fall. Kein Institut lässt sich freiwillig evaluieren und hat nur wenig Einfluss auf das Evaluationsverfahren, denn es wird turnusmäßig von Wissenschaftsorganisationen zur Teilnahme aufgefordert. Alle Evaluationen stellen damit ohne akuten Anlass den inhaltlichen und finanziellen Fortbestand von Instituten in Frage und untermauern diesen Anspruch mit der Kopplung an Förderentscheidungen. Unter diesen Vorzeichen sind Evaluationen aus Sicht der Institute künstlich erzeugte Krisensituationen, die es wie eine Prüfung möglichst unbeschadet zu überwinden gilt. Empirisch drückt sich das zum einen in der selbstverständlichen Orientierung an einer positiven und problemlosen Evaluierung aus. Zum anderen sind die umfangreichen Vorbereitungshandlungen der Institute zu nennen, die sämtlich darauf abzielen, bereits im Vorfeld alles zu tun, um den Ruf, die Finanzierung und die inhaltliche Ausrichtung der jeweiligen wissenschaftlichen Einrichtung nicht zu gefährden. Das

kämpfen. Vgl. den Beitrag von Thamar Klein und Meike Olbrecht in diesem Band sowie Lamont 2009, Langfeldt 2001, Travis/Collins 1991.

ideale Ziel ist deshalb, von den Gutachter/innen möglichst nicht kritisiert, sondern unterstützt zu werden. Folglich muss man die Vorbereitungshandlungen der Institute als Überzeugungs- oder Kritikvermeidungspraktiken interpretieren, die zu einer Bestätigung führen sollen und je nach Verfahrenstypus variieren. Sie reichen von der Schaffung von Unterstützungsstrukturen zur Produktion von Forschungsoutputs, der Selektion geeigneter Kandidaten für die Präsentationen vor Ort oder der Einreichung von Outputs über das Einkaufen von Stars bis zum aufwendig betriebenen Eventmanagement mit Probeevaluationen. Bei aller Verschiedenheit dieser Praktiken haben sie doch alle das gleiche Ziel: Alles was die Gutachtenden beobachten könnten, wird vorab einer intensiven internen Überprüfung unterzogen. Die wissenschaftsspezifische Präferenz für eine kritische Auseinandersetzung ist in eine öffentlich einsehbare Handlungssituation eingebettet, die für evaluierte Kollegen potentiell existenzgefährdend ist. Ob daraus eine Tendenz zur wenig kritischen Begutachtung im Sinne eines „Nichtangriffpakts" (Schimank 2005: 149) beziehungsweise ein bloßes „akademisches Ritual" (Michaels 2010) folgt, Gutachter/innen ihre Position als „epistemic elites" nutzen und als „arbiters of excellence" partikulare Maßstäbe durchsetzen (Martin/Whitley 2010: 67) oder vielleicht doch gemeinsame Grundorientierungen bei der Urteilsfindung zu beobachten sind, diskutieren wir im Abschnitt 5.2.

Eine dritte Verschiebung betrifft schließlich den *sozialen Kontext der Urteilsfindung* selbst. Statt eines individuellen und anonymen Gutachtervotums steht jetzt ein öffentlich zugängliches und kollektiv getragenes Urteil eines zumeist heterogen zusammengesetzten Gutachterteams im Zentrum. Deshalb stellt sich die Frage, ob das Gutachterurteil durch spezifische Gruppendynamiken präformiert wird und welcher Art diese sind. Stößt man auf Kompromissbildungen zwischen unterschiedlichen Positionen? Reihen sich nur Einzelmeinungen aneinander? Erzeugen erst Verfahrensregeln oder Outputquoten eine Einigung? Oder finden in solchen Gutachtergruppen Normbildungsprozesse auf Grundlage von geteilten Standards statt? (siehe 5.3).

Institutionelle Evaluationen bilden also in mehrerer Hinsicht eine neuartige und spannungsreiche Handlungssituation für die Gutachter/innen. Wie sie auf die Aufforderung an diesen teilzunehmen reagieren, welche Grundorientierungen dabei sichtbar werden und auf welche Weise sie diese neue Herausforderung bewältigen, erörtere ich im Folgenden.

4 Reaktionsweisen und Grundorientierung von Gutachter/innen: Vom wissenschaftspolitischen Auftrag zur professionellen Verpflichtung

Wissenschaftlerinnen und Wissenschaftler reagieren nicht passiv auf neuartige Regulierungsformen, sondern bringen ihre eigenen Orientierungen aktiv ein. Die Frage, warum und wozu Gutachter/innen an in mehrerer Hinsicht spannungsreichen und sehr arbeitsintensiven Evaluationen teilnehmen, ist ein instruktives Beispiel hierfür. Gutachter/innen werden für gewöhnlich von den jeweiligen Evaluationsagenturen angefragt und überlegen nicht lange, ob sie teilnehmen sollen. Sofern die Zeit es zulässt, gilt ihnen die Teilnahme als eine Selbstverständlichkeit.[7] An Evaluationen zu partizipieren wird weder als ein grundsätzlich problematisches noch ablehnbares Unterfangen thematisiert. Die Teilnahme gilt unter den Gutachterinnen und Gutachtern als eine Pflicht gegenüber der scientific community, der man sich nicht entziehen darf: „I think part of my responsibility is not only to do my job in isolation, but to do my job in the context of the community that I relate to, and it's in the interest of that community". Sich in den Dienst der jeweiligen Gemeinschaft zu stellen heißt natürlich nicht, keinen eigenen Vorteil hieraus zu ziehen. So fühlen sich die Gutachter/innen geehrt, diese Rolle zu übernehmen, sie interessieren sich für das Innenleben der evaluierten Einrichtung, können einen Überblick über die Entwicklung und den aktuellen Stand von Forschungsgebieten bekommen,

7 Diese Selbstverständlichkeit tritt in den Interviews dergestalt auf, dass die Gutachter/innen auf die Frage „Wie kam es dazu, dass Sie Gutachter/in wurden?" allenfalls spekulierten, warum Evaluationsagenturen sie angefragt hatten. Ihre Eigenmotivationen blieben dabei aber ausgeblendet und mussten explizit erfragt werden.

sie sind auf die Sicht- und Begründungsweisen der Gutachterkollegen gespannt, können über den Ablauf solcher Evaluationen etwas für die zukünftige Beurteilung der eigenen Institution lernen und schließlich muss die professionelle Selbstkontrolle gewahrt bleiben: „I think it's, since somebody has to do it, it's better that […] you don't refuse to participate". Die Verfolgung genuin wissenschaftlicher Eigeninteressen, die Gewissheit auch über ganze wissenschaftliche Einrichtungen ein Urteil fällen zu können[8] und die Verpflichtung zur ‚community work' setzen Evaluationen in ein Licht, als wären sie eine gewohnte wissenschaftliche Praxis. Ein Ausdruck davon ist, dass die Gutachter/innen die konkreten Evaluationskriterien der Verfahren oftmals nicht präsent haben. Wissenschaftspolitisch initiierte Evaluationen werden also durch die Brille von Wissenschaftler/innen beobachtet, entlang der dort gültigen Regeln, Normen und Wertorientierungen interpretiert und letztlich durch das Einrücken in ihre Handlungsroutinen normalisiert.

Das zeigt sich auch an der spezifischen Interpretation von Gutachter/innen, wozu und für wen Evaluationen dienlich sind. Besonders auffällig ist nämlich die Grundhaltung, Evaluationen weniger als eine Außenkontrolle im Dienst der Wissenschaftspolitik denn als *kollegiale Unterstützung* aufzufassen. Selbst im hierzu weniger geeigneten RAE, das ja einen numerischen Verteilungsschlüssel und keine inhaltlichen Rückmeldungen erarbeitet, findet sich eine solche Primärorientierung: „The first reason (for the RAE) which is probably the only valid one, really is *to help* universities benchmark their research against their competitors". In den beiden anderen Verfahren, tritt sogar ein Selbstverständnis der Gutachtenden als *kollegial beratende Instanz* deutlich hervor: „We are not sitting there to make difficulties, to look for all that's wrong". „It was not as much to find faults but, are they going in the right direction? And are there points of improvement? Can things be done in a better way?"

Die Evaluationsverfahren mit ihren interpretationsbedürftigen Kriterien („Qualität", „Produktivität", „Effektivität", etc.) und die Positionierung von Fachgutachtenden an zentraler Entscheidungsstelle bieten zwar den Raum, professionseigenen Normen und Handlungsweisen zu folgen, sie ziehen aber auch Grenzen. Das kann man gut an den Begutachtungspraktiken beider Evaluationstypen beobachten.

5 Begutachtungspraktiken

Ich konzentriere mich im Folgenden auf ausgewählte Aspekte des Gutachterhandelns, die unmittelbar mit den zuvor angeführten Verschiebungen im Rahmen institutioneller Evaluationen verknüpft sind. Gutachter/innen sind keine Organisationsanalysten und müssen dennoch über die Leistung(sfähigkeit) von wissenschaftlichen Einrichtungen urteilen. Mit welchen Selektivitäten machen sie sich diesen neuen Gegenstand handhabbar? (5.1) Gutachter/innen müssen unter großem Handlungsdruck potentiell folgenreiche Entscheidungen über ihre Kollegen fällen. Welche Standards kommen dabei zum Einsatz und inwiefern genügen sie wissenschaftlichen Gütekriterien? (5.2) Gutachter/innen mit je eigenen Sichtweisen treffen in Panels kollektive Entscheidungen. Unterliegen solche Gruppenentscheidungen speziellen Selektivitäten oder werden nur Einzelmeinungen von Gutachtern hintereinander gestellt? (5.3)

5.1 Handhabung des neuen Gegenstands ‚wissenschaftliche Organisation'

Alle Verfahren fordern von den evaluierten Einrichtungen heterogene Informationen an. Den Gutachtergruppen bleibt aber überlassen, welchen Stellenwert verschiedene Informationsquellen bei der Urteilsfindung haben sollen. Die Verfahren liefern eine Vielzahl möglicher Betrachtungsweisen, aus denen die Gutachter/innen die fach- und fallspezifisch angemessenen wählen. Verfolgt man, welche Informationen die Gutachter/innen aus welchen Gründen selektieren, dann sind die Zuverlässigkeit

8 Eva Barlösius (2008) spricht von einer „Urteilsgewissheit" von Gutachterinnen und Gutachtern selbst bei Fragen, die über ihre Fachexpertise weit hinausreichen.

sowie die Kompatibilität mit eingeübten Begutachtungsweisen und den primär verfolgten Evaluationszielen von zentraler Bedeutung.

Das britische Research Assessment Exercise gewichtet bereits qua Verfahren die drei zentralen Informationsquellen. Das Gesamturteil soll mindestens zu 50% auf „outputs" und jeweils zu 5% auf „environment"- bzw. „esteem"-Indikatoren beruhen. In der Begutachtungspraxis lässt sich eine weitere Konzentration auf den Stellenwert von „outputs" beobachten (zwischen 60 % bei den Ingenieuren und 80% bei den Historikern). Der Gesamtfall ‚Organisation' wird bereits durch dieses Verfahren in Fachbereiche und Einzelinformationen zerlegt und dann in publizierte Einzelakte weiter klein gearbeitet. Erst am Ende werden die Einzelurteile (unqualified, Noten von 1 bis 4) addiert und mit den ebenfalls einzeln bewerteten „esteem"- und „environment"-Informationen zu einem numerischen „quality profile" verrechnet. „Esteem" und „environment" treten in den Hintergrund, weil diese Informationen unter den Gutachter/innen als schwer kontrollier- und einschätzbare Selbstdarstellungen gelten. Sie dienen ihnen jedoch als ein flexibel einsetzbares Korrektiv, z.B. um dem Problem zu begegnen, dass Publikationen nicht in allen Disziplinen den gleichen Stellenwert haben. Im RAE wird der komplexe Gegenstand „wissenschaftliche Einrichtung" durch die Zerlegung in einzelne Fachgebiete, in Einzelleistungen und Einzelurteile handhabbar gemacht. Auf diese Weise bleibt der Eigenwert organisatorischer Fragen begrenzt und die Gutachtenden können auf ihre Erfahrungen aus dem klassischen Peer Review zurückgreifen.

Eine ganz andere Selektivität findet man hingegen bei den niederländischen und deutschen Verfahren. Dort sollen fallspezifische Urteile und Empfehlungen bezüglich einer Gesamtorganisation gefällt werden. Hierfür muss man jedoch den jeweiligen Fall aus verschiedenen Einzelinformationen zunächst konstruieren und zu einem Gesamtbild zusammenfügen und nicht wie im RAE in Einzelakte zerlegen: „You have an impression and you try to articulate that impression and then those criteria are helpful". Deshalb richtet sich die Aufmerksamkeit der Gutachter/innen vor der Begehung insbesondere auf die Programmatiken, Selbstevaluationsberichte und Stärken-Schwächen-Analysen der Gesamtorganisation und Forschungseinheiten. Vor allem dort wird nämlich das Selbstbild der Institution konstruiert, das unter Zuhilfenahme weiterer Informationsquellen von den Gutachter/innen auf seine Konsistenz und Tragfähigkeit überprüft wird. So gibt der Vergleich mit dem vorherigen Evaluationsbericht Aufschluss über die Problemwahrnehmung und Problemlösungskapazität der Institution. Auch die umfangreichen Publikationslisten, (Drittmittel-)Bilanzen oder Zitationsanalysen dienen den Gutachtern nicht so sehr für die direkte Bewertung des Gesamtfalls.[9] Vielmehr sind es flankierende Informationen, die relativ unabhängig von den Selbstbeschreibungen sind und insofern ein eigenständiges Gesamtbild zeichnen, das dann zur Konsistenzprüfung eingesetzt werden kann. Die Glaubwürdigkeit und Angemessenheit der aus Gutachtersicht zentralen „Selbstbeschreibungen mit dem üblichen Selbstlob" werden also ebenso wenig wie im RAE einfach vorausgesetzt. Aber diese Verfahren bieten die Möglichkeit, den Realitätsgehalt und die Angemessenheit solcher Selbstdarstellungen spätestens während der interaktiven Begehungen zu überprüfen und so als analytisches Mittel zu nutzen. Die Gutachter/innen nehmen in diesen Verfahren also tatsächlich die gesamte Einrichtung in den Blick und müssen hierfür ein Gesamtbild generieren. Es steht deshalb der Zusammenhang von einzelnen Informationen im Zentrum und nicht wie im RAE die Zerlegung in einzelne Aspekte. Mit der Konsistenzprüfung von Programmatiken stellen die Gutachter/innen auch hier Informationen ins Zentrum, die sie durchaus gewohnt sind zu bewerten und zu kommentieren. Sie behandeln den Gegenstand Organisation entlang den Inhalten wissenschaftlich relevanter Produkte.

9 Aus der Perspektive der evaluierten Institute mag schon deshalb der Eindruck einer herausragenden Bedeutung der quantitativen Leistungsbemessung entstehen, weil ein großer Teil der Vorbereitung in der Erstellung von Tabellen und Etablierung von Monitoring-Systemen besteht. Dass die Gutachter/innen maßgeblich auf Grundlage dieser Daten urteilen, kann auf Grundlage unserer Erhebung jedenfalls nicht bestätigt werden. Diese quantitativen Informationen helfen schließlich nur wenig in einem Verfahren, das in hohem Maße auf Empfehlungen hinsichtlich der zukünftigen Organisationsentwicklung ausgelegt ist.

5.2 Begutachtungsweisen im Kontext der Verfahren

Wie gehen die Gutachter/innen aber mit den jeweils ausgewählten Primärquellen um und welche Grundorientierungen bringen sich darin zum Ausdruck? Aus der Pflicht gegenüber der scientific community folgt zunächst, dass Gutachter/innen ihre Aufgabe sehr ernst nehmen und einer zweipoligen Haltung folgen: „to make sure we did *the job properly* but that we were also *very fair to the community*". Damit sind einerseits die Einhaltung akzeptabler Standards („proper job") und andererseits die angemessene Anwendung auf die jeweilige Wissenschaftsart („fair job") gemeint.[10] Diese Haltung stellt ein Regulativ dar, damit der Begutachtungsprozess weder in eine unkritische Interessenpoltik für das eigene Fachgebiet noch in überkritische Leistungsanforderungen abgleitet.[11] Die Durchsetzung dieser Grundhaltung trifft allerdings auf verfahrensspezifische Problemlagen.

Mit der Fokussierung auf einzelne „outputs" im Research Assessment Exercise entsteht für die Gutachter/innen zunächst das Problem, viele hundert Einzelbegutachtungen vornehmen zu müssen: „The major challenge for quality was getting the job done!" Deshalb ist auch erwartbar, dass unter Handlungsdruck Beschleunigungs- und Rationalisierungsstrategien entwickelt werden. Aufschlussreich ist dabei, wie abgekürzt wird und aus welchen Gründen heraus. Es wäre naheliegend und im Rahmen des Verfahrens auch möglich, dass die Gutachter/innen ihre Arbeitslast durch die Auswahl nur weniger Outputs, den Verzicht auf eine (zumindest) von zwei Gutachter/innen vorgenommene Bewertung oder durch die Hinzunahme von metrischen Daten bereits publizierter und begutachteter Outputs bewältigten. Genau das geschieht aber nicht, weil die Gutachter/innen auf diese Weise unzulässig abkürzen würden und kein gesichertes Urteil fällen könnten.[12] Die Beschleunigung und Rationalisierung der Einzelbegutachtungen erfolgt vielmehr durch eine Radikalisierung von wissenschaftsspezifischen Deutungsschemata. Das Spezielle solcher Deutungsschemata ist, dass es sich um hochgradig implizite und generalisierte Gesichtspunkte handelt, mit deren Hilfe Outputs durchmustert werden. Ein beschleunigtes Lesen und Interpretieren ist hierfür zwingend erforderlich. Den impliziten Charakter dieser Deutungsschemata kann man sich an dem Phänomen vergegenwärtigen, dass die Gutachter/innen nicht von einhelligen Begutachtungsweisen im Team ausgingen und entsprechend überrascht waren, als sie in sogenannten „calibration sessions" und den nachfolgenden Einzelbewertungen auf eine hohe Übereinstimmung der Urteilsweisen und Urteile stießen. In diesen „calibration sessions" haben die Gutachter/innen ihre impliziten Urteilsweisen explizit gemacht. Dafür hat jeder Publikationen ausgewählt, die aus individueller Sicht das Notenspektrum zwischen ‚unclassified' und der Bestnote 4 repräsentieren. Dann haben alle Panelmitglieder unabhängig voneinander dieses Sample bewertet. Neben der Erkenntnis erstaunlich ähnlicher Urteilsweisen wurde bei der Diskussion von abweichenden Fällen deutlich warum und wer zu harsch oder zu milde urteilt. Die Urteilsweisen von Gutachterinnen und Gutachtern wurden so ausfindig gemacht und eingeordnet: „Very quickly you create an ethos

10 Michèle Lamont et al. (2009a) sprechen von „Fairness as Appropriateness". Damit ist vor allem eine „epistemic contextualisation" gemeint. Diese Norm sorgt dafür, dass epistemologische Vorlieben nicht einfach auf jeden beliebigen Fall projiziert werden, sondern der zu begutachtende Gegenstand über die Angemessenheit von Standards entscheidet.

11 Die Geltung der Norm zeigt sich besonders gut bei Abweichungen. Zum Beispiel wurde von einem schwierigen Gutachterkollegen berichtet, der eine neue Forschungsrichtung vertrat, generell bevorzugte und in seinen Urteilen entsprechend von der Gutachtergruppe abwich. Der Vorsitzende suchte das Gespräch und dieser Gutachter revidierte seine Urteile. Das klassische Beispiel für überzogene Leistungserwartungen ist die Frage danach, was ein wirklich sehr guter und relevanter Beitrag ist. „Paradigm shifting work" und selbst „interesting ideas which change your look on a certain problem" sind nicht erwartbar und gerade in einer Situation kein geeigneter Maßstab, in der eine überkritische Bewertung zu existentiellen Problem führen kann.

12 Interessanterweise gibt es unter den Gutachterinnen und Gutachtern immer wieder den Verdacht, dass zwar nicht im eigenen, aber sicherlich in anderen Panels so vorgegangen würde. Besonders deutlich ist die Ablehnung illegitimer Abkürzungen hinsichtlich einer rein bibliometrischen Bewertung von Outputs, die alle Gutachter/innen aus den bekannten Gründen ablehnen: Weder die Häufigkeit der Zitationen, die Menge oder der Publikationsort gebe Aufschluss darüber, ob es sich im konkreten Fall um einen guten Forschungsbeitrag handele.

for the panel". An den Standards ist interessant, dass sie weder absoluten Maßstäben folgen noch vom jeweiligen Begutachtungsfall ablösbar sind. So könne die Bestnote eigentlich nur an Outputs mit dem Potential zum Paradigmenwechsel – also fast nie! – vergeben werden, so dass es sich stets um ein relatives Urteil über realistisch erwartbare Leistungen handelt. Ebenso ist es unmöglich ein sachhaltiges Urteil zu fällen, wenn man nicht die konkrete Publikation in Augenschein nimmt. Ein beliebtes Beispiel hierfür sind Reviews. Sie können eine einfache Zusammenfassung aktueller Forschungen und damit zwar nützlich, aber ohne wissenschaftlichen Eigenwert sein, selbst wenn sie im Topmagazin Nature veröffentlicht und viel zitiert werden. Umgekehrt ist es aber auch möglich, dass in dem Reviewartikel eine neuartige Frage oder Erklärung generiert wurde und dieser damit äußerst wertvoll ist. Das lässt sich aber nur durch Lesen und Deuten herausfinden.

Ist die grundsätzliche Frage geklärt, ob es sich überhaupt um einen eigenständigen Forschungsbeitrag handelt, dann wechselt die Aufmerksamkeit der Gutachtenden von der Textgattung auf die Textgestalt. Entlang wichtiger Deutungsdimensionen explorieren sie an geeigneten Textstellen die Werthaltigkeit des Beitrags: Was ist der vom Text selbst gesetzte Anspruch? (Überschrift) Was ist daran neu? (Abstract) Stützt die empirische Evidenz den erhobenen Anspruch? (Datengrundlage) Ist die Argumentation klar und konsistent? Welchen verallgemeinerbaren Wert haben die empirischen Einzelergebnisse? (Diskussion). Gelingt dem Text keine gelungene Kommunikation entlang dieser Dimensionen, dann schließen Gutachter/innen auf Schwächen der unternommenen Forschung. Der Begutachtungsprozess bezieht sich also auf den zur Verfügung stehenden Gegenstand, d.h. vor allem auf die Art und Weise wie Forschung kommuniziert wird: „It's all about communication". Durch die Verwendung textbezogener Relevanz, Konsistenz und Stimmigkeitskriterien kann die Begutachtung beschleunigt werden. Eine ganz andere Beschleunigungsform folgt schließlich aus dem Verfahren selbst. Bei den wenigen stark abweichenden Voten ist die Bereitschaft zur Angleichung der vergebenen Noten schon deshalb groß, weil die Bewertung eines einzelnen Outputs für die kumulative Gesamtnote kaum ins Gewicht fällt. Das Gutachterhandeln ist also weiterhin durch wissenschaftliche Normen bestimmt, deren Einhaltung im Kontext eines aufwendigen Verfahrens allerdings immer schwieriger wird. Diese Normen unterbinden (bislang) vielleicht effizientere, aber als illegitim angesehene Abkürzungsstrategien und erzeugen erst den von allen Gutachtern beklagten Zeitaufwand.[13]

Auch bei den niederländischen und deutschen Verfahren spielen allgemeine Deutungsschemata der Stimmigkeit eine wichtige Rolle. Vor allem jedoch auf einer höher aggregierten Ebene als im RAE: Statt Einzelleistungen stehen Gesamtprogrammatiken im Zentrum. Widersprüchlichkeiten, Inkonsistenzen oder fehlende Explikationen geben den Gutachtenden den Anlass für Nachfragen, Kritiken und den Verdacht, dass die Autoren selbst nicht genau wüßten, wofür sie was tun. Sowohl beim Lesen der Unterlagen wie auch der interaktiven Begehung sind die Gutachter/innen auf der Suche nach möglichen Problemlagen in der Ausrichtung und den Arbeitszusammenhängen des Instituts. Das ist natürlich in einem prüfungsartigen Evaluationskontext besonders schwierig, weil dort ja gerade nicht von einer freimütigen Offenlegung von bearbeitungswürdigen Problemlagen ausgegangen werden kann. Die Darstellung der Erfolge steht für die Institute im Vordergrund und führt gelegentlich zu einem institutsinternen Wettbewerb darüber, wer präsentieren darf.

Die Gutachter/innen sind darauf gefasst, hinter die mit allerlei rhetorischen, ästhetischen und inszenatorischen Mitteln aufgebaute Fassade solcher Selbstbeschreibungen gelangen zu müssen, um urteilen zu können. Die Stärken-Schwächen-Analysen und die Umgangsweise mit den Empfehlungen der letzten Evaluation bieten den Gutachterinnen und Gutachtern einen Zugang, um einen ersten Einblick in die Problemsicht und Problembearbeitungsweisen der jeweiligen Einrichtung zu bekommen. Sie begeben sich aber noch weiter auf Indiziensuche und verwenden im Wesentlichen

13 Im nächsten RAE, Research Excellence Framework (REF) genannt, soll die Verwendung von metrischen Daten wichtiger werden. Ob die Gutachter/innen trotz massiver Kritik aus pragmatischen Gründen dennoch auf sie zurückgreifen, bleibt abzuwarten.

drei Strategien, um die sachliche Angemessenheit, Glaubwürdigkeit und Realisierbarkeit von Institutsprogrammatiken zu überprüfen. Erstens werden die Selbstdarstellungen an zusätzlichen Informationen gespiegelt, die teilweise zur Verfügung gestellt werden (Outputs, Infrastruktur, Finanzierung, Kooperationsbeziehungen), vor allem aber den Gutachter/innen „auch ohne die Unterlagen" vorliegen, weil sie auf ein „gewisses Vorwissen" rekurrieren können. Dieses Vorwissen ist ein Erfahrungswissen, das zum Beispiel die Besonderheiten des wissenschaftlichen Themenbereichs, die institutionelle Einbettung des zu begutachtenden Instituts oder typische Schwierigkeiten des Wissenschaftsbetriebs umschließt. Es handelt sich dabei also weniger um ein spezialisiertes und formalisiertes Ableitungswissen, als viel mehr um eine generelle Kenntnis des Handlungsfeldes mit seinen typischen Herausforderungen und Problemlagen: „Irgendwie weiß man das".[14] Zweitens erfolgt eine textimmanente Konsistenzprüfung: „Die Papiere müssen in sich logisch sein oder ne gewisse, einen Zusammenhang aufweisen". Bereits im Studium der Akten suchen die Gutachtenden neben Widersprüchlichkeiten auch nach Indizien für und wider die Glaubwürdigkeit der Selbstdarstellungen, die dann in der direkten Interaktion eine weitere Überprüfung erfahren. Denn „die Papiere [müssen] mit dem was die Leute sagen in Übereinstimmung stehen, so dass es nicht auseinanderfällt". Die Gutachter/innen kommen also bereits mit einem ersten Bild im Kopf in die Institute, das dort bestätigt, verfeinert, ausgebaut oder relativiert, aber nach eigener Auskunft nur selten verworfen wird.

Eine dritte Strategie besteht schließlich darin, die spontanen Reaktionsweisen (insbesondere der Institutsleitung) auf Fragen genau zu beobachten und Schlussfolgerungen daraus zu ziehen. Wie in einer therapeutischen Situation oder in juristischen Verfahren kommt es dabei weniger darauf an, was den jeweiligen Fall betreffend gefragt wird, sondern wie die Antworten ausfallen. Hierzu ein Gutachter:

> *„Also das ist eben die Art, antworten die Leute auf Fragen, die man ihnen stellt, auch auf kritische Fragen, und wie gehen sie mit diesen Fragen um? Weichen sie denen aus, beantworten sie die gar nicht, beantworten sie die klar? Wenn sie die klar beantworten und auch ein Problem eingestehen, ist das im Prinzip schon mal ein Indiz, dass es in die richtige Richtung geht. Und wenn sie ein realistisches Selbstbild auch haben, wie sie sich selber einschätzen, ist das auch ein positives Indiz. Also das sind Indizien, die was mit Glaubwürdigkeit zu tun haben."*

Die Handlungsweise der Gutachter/innen geht weit über das Muster fachlicher Expertise und Kritik hinaus. Vor allem fällt die Nähe zu dem auf, was man in der Soziologie den Bereich professionellen Handelns nennt und in diesem Beispiel die Form des beratenden oder supervisorischen Handelns annimmt. Die Bewertung steht nicht für sich, sondern soll Verbesserungen anregen. Dafür ist aber konstitutiv, dass Probleme offengelegt und kommuniziert werden. Erst vor diesem Hintergrund wird verständlich, dass Gutachter/innen das Eingestehen von Problemen, ein realistisches und problembewusstes Selbstbild der Institute, honorieren, obwohl in der Handlungssituation Evaluation gerade nicht damit zu rechnen ist. Das kritische kollegiale Gespräch über Problemlagen und mögliche Lösungen ist eine wissenschaftsintern besonders anschlussfähige Deutung von Evaluationen, der aber eine verfahrensbezogene Deutung als Prüfung entgegensteht.[15]

An beiden Beispielen von Gutachterpraktiken im Kontext institutioneller Evaluationen müsste deutlich geworden sein, dass Eigenregulierungen der Wissenschaft nicht einfach ausgehebelt werden und noch immer das Handeln strukturieren. Die Verfahren erschweren dies allerdings in unterschiedlicher Weise. Sie generieren einen Zeitdruck, erzwingen Abkürzungsstrategien oder erzeugen eine Prü-

14 Im juristischen oder ärztlichen Handeln entspricht dem ein typologisches Wissen über Standardfälle, Symptomatiken, Problem- oder Motivkonstellationen, mit denen ein konkreter Fall kontrastiert und in seiner Spezifik erschlossen werden kann.

15 So tituliert ein Institutsdirektor seinen Erfahrungsbericht mit der Evaluation der Leibnizgemeinschaft mit „Die Prüfung als Chance begreifen" und betont den „Dialog-Charakter" gegenüber der „Kontrollvisite". Vgl. Leibniz-Journal 3/4, 2006, S. 30f.

fungssituation, die der Verständigung über Verbesserungsmöglichkeiten entgegen steht.

5.3 Gruppendynamik

Kommen wir zur letzten Verschiebung der Handlungssituation von Gutachterinnen und Gutachtern im Kontext institutioneller Evaluationen. Im Unterschied zum klassischen Peer Review fällen die Gutachter/innen ihr Urteil in einer Gruppe. Genauer gesagt müssen Einzelurteile, die Gutachter/innen bei der Durchsicht von Unterlagen, Outputs oder bei der Begehung gefällt haben, von der gesamten Gruppe getragen und gemeinsam nach Außen vertreten werden. Heterogen besetzte Gutachtergruppen stehen also unter einem erhöhten Konsenszwang. Dieses Problem wird in den Verfahren unterschiedlich bewältigt. Im RAE erfolgt eine Abstimmung der Bewertungsweise in den bereits genannten „callibration sessions" vor den Einzelbegutachtungen, es kommt zu Diskussionen zwischen den (meist) zwei für einen Output verantwortlichen Gutachtern, sofern die Urteile zu weit auseinander gehen, und es gibt Konsistenzkontrollen der Urteilsweisen von Einzelgutachtern während des Prozesses. Diese bestehen darin, dass die Varianz der von Gutachtern vergebenen Noten überprüft und ggf. korrigiert wird (siehe das Beispiel in FN 11). Bei der Bildung des Gesamtergebnisses bedarf es keiner weiteren Koordination, da die Einzelnoten nur zusammengerechnet werden. In den niederländischen und deutschen Verfahren werden die Einzeleindrücke der Gutachter/innen im Verlauf der Begehung immer wieder in Gesprächen der Gutachtergruppe gesammelt. Besonders wichtig ist das erste Treffen am Vorabend der Begehung und das letzte, bevor eine Rückmeldung an die evaluierte Einrichtung gegeben wird. Die erste Ab-stimmung gibt vor, ob es sich nach Lektüre der Unterlagen um eine problematische oder unproblematische Evaluation handelt. In der letzten wird selektiert, welche Kritikpunkte und konkreten Empfehlungen von der Gutachtergruppe letztlich vertreten werden. Oftmals werden die als Checklisten vorliegenden Kriterienkataloge erst hier herangezogen, um die gesammelten Eindrücke zumindest im Bewertungsbericht ans Verfahren anzugleichen.

In beiden Evaluationstypen finden sich also Hinweise darauf, dass Gruppenurteile nicht auf das Hintereinanderschalten von Einzelperspektiven der Gutachter/innen reduziert werden können. Beschreibungen wie „ethos for the panel" und der Rückgang auf sehr allgemeine Deutungsschemata verweisen vielmehr darauf, dass in diesen Gruppen grundsätzliche Normen und Wertorientierungen wissenschaftlichen Handelns aktualisiert, eingefordert oder sogar erst gebildet werden, um über Sonderperspektiven hinaus zu gelangen. Die Orientierung an abstrakten und kaum operationalisierbaren Standards macht das Gutachterhandeln zwar wenig berechenbar, aber dennoch zweckrational: Sie ermöglichen eine *Beschleunigung der Urteilsfindung*, weil eine Reduktion auf wesentliche Aspekte erfolgt und nur strittige Sachverhalte diskutiert werden. Die Notwendigkeit, Einwände gegenüber wissenschaftlichen Kollegen begründen und rechtfertigen zu müssen, hat zudem eine *disziplinierende Wirkung*. Denn jeder Gutachtende wird in diesen Gruppen zugleich selbst begutachtet.[16] Zu hart oder zu milde Urteilende, zurückhaltende oder viel diskutierende, sach- oder selbstbezogene Gutachter/innen werden sichtbar. Daran schließen *diskursive Effekte* an, denn der Wert dieser Beiträge bemisst sich daran, ob sie Anschlussfähigkeit in der Gruppe erlangen. Kritische Einwände oder Sonderperspektiven müssen sich nämlich in der Gutachtergruppe bewähren und werden nicht einfach übernommen. Sie finden ihren Weg in das Gesamturteil nur, wenn sie von anderen Gutachter/innen aufgegriffen und damit bekräftigt werden.[17]

16 Im „Forum: Begutachtung in der soziologischen Drittmittelforschung" auf dem Kongress der Deutschen Gesellschaft für Soziologie 2010 haben die sog. Fachkollegiaten, die in ihrem Gremium ebenfalls Gruppenentscheidungen treffen, explizit erwähnt, dass sie ihre Funktion in der Begutachtung der Gutachter/innen sehen und dieser vornehmlich durch Normbildungsprozesse innerhalb der Gruppe hinsichtlich der Angemessenheit von Gutachten nachkommen.

17 Das kann dann auch dazu führen, dass in Begutachtungen ein vergleichsweise singulärer Aspekt herausgegriffen wird, weil dieser allen Gutachtern aufgefallen ist. Sofern ein Fachexperte unter den Gutachtenden ist, hat dieser erhöhte Durchsetzungschancen, weil dieser den Wert eines Beitrags seinen Gutachterkollegen besser begründen

Im Anschluß an Michèle Lamonts These, dass bei Gruppenevaluationen kaum formalisierbare „Customary Rules“ (2009) und die fallbezogene „Appropriateness“ (2009a) des Urteils eine zentrale Bedeutung haben, habe ich zu zeigen versucht, dass sehr allgemeine und nicht mechanistisch anwendbare Deutungsschemata der Konsistenz, Stimmigkeit und Fallangemessenheit auch unter den spezifischen Rahmenbedingungen institutioneller Evaluationen operieren. Diese scheinen gerade in heterogenen Gremien sogar eher wichtiger zu werden und sind sicherlich nicht durch die formalen Verfahrensregeln erzeugt worden.

6 Fazit

Was kann man aus unserem Ausflug in die Innenwelt von Evaluationsprozessen über die institutionellen Folgen von Evaluationen lernen? Ich habe an einer rein institutionalistischen Perspektive kritisiert, dass formale Rahmenbedingungen nicht einfach das empirisch beobachtbare Handeln von Gutachterinnen und Gutachtern bestimmen. Diese agieren auf Grundlage eigener Regeln der Angemessenheit, die keineswegs vom formalen Verfahren erzeugt wurden. Welche Zwecke Verfahren der institutionellen Evaluation auch immer verfolgen, sie sind in ihrer Durchführung zumindest so lange mit wissenschaftsinternen Relevanzen durchdrungen, wie Angehörige der scientific community an zentraler Entscheidungsposition platziert sind. Vor allem geben solche Verfahren an, was begutachtet und in welcher Form Ergebnisse kommuniziert (numerisch/inhaltlich) werden sollen. Wie Gutachter/innen begutachten sollen bleibt jedoch unbestimmt, so dass sie hinreichend Freiraum haben, die Verfahren an wissenschaftliche Standards anschlussfähig zu halten.

Das habe ich an drei wesentlichen Neuerungen, die institutionelle Evaluationen mit sich bringen, gezeigt. Auch wenn dort wissenschaftliche Organisationen beurteilt werden, so stehen mit der Konsistenz und Innovativität von Forschungsprogrammen oder Forschungserzeugnissen weiterhin wissenschaftsspezifische Aspekte im Zentrum. Organisation wird gewissermaßen auf die Frage begrenzt, ob vorhandene Regelungen geeignet sind, interessante Forschung zu unterstützen. Mit diesem neuen Gegenstand geht allerdings das Problem einher, dass eine Vielzahl heterogener Informationen von den Gutachtern durchgemustert werden müssen und ihr Urteil für die Zukunft einer ganzen wissenschaftlichen Einrichtung folgenreich sein kann. Gutachter/innen stehen damit unter erhöhtem Handlungsdruck, tragen große Verantwortung und haben einen erheblichen Einfluss. Die gerne daraus abgeleiteten Tendenzen zu schematischen, milden oder partikularistischen Urteilsweisen konnten allerdings nicht bestätigt werden. Vielmehr stößt man unter den Gutachtern auf eine Haltung, kollegiale Unterstützung über eine Fall angemessene Kritik auszuüben, die sich an abstrakten Fächer- oder Gegenstandsgrenzen übergreifenden Interpretationsschemata orientiert. Diese sind gerade in heterogen besetzten Gutachtergruppen ein wichtiges Mittel, um zu einem gemeinsamen Urteil zu gelangen. Die paradoxe Folge extern initiierter Überprüfungen im Rahmen institutioneller Evaluationen könnte also sein, dass wissenschaftseigene Kriterien radikalisiert und auf neue Gegenstandsbereiche ausgedehnt werden.

kann.

Literatur

Barlösius, Eva, 2008: Urteilsgewissheit und wissenschaftliches Kapital, in: *Matthies, Hildegard/ Simon, Dagmar (Hg.):* Wissenschaft unter Beobachtung: Effekte und Defekte von Evaluationen. Wiesbaden: VS Verlag, 149-196.

Gläser, Jochen / Lange, Stefan / Laudel, Grit / Schimank, Uwe, 2008: Evaluationsbasierte Forschungsfinanzierung und ihre Folgen, in: *Mayntz, Renate et al. (Hg.):* Wissensproduktion und Wissenstransfer. Wissen im Spannungsfeld von Wissenschaft, Politik und Öffentlichkeit. Bielefeld: Transcript, 145-170.

Gläser, Jochen / Laudel, Grit, 2007: Evaluation without Evaluators: The impact of funding formulae on Australian University Research, in: *Whitley, Richard / Gläser, Jochen (eds.):* The Changing Governance of the Sciences: The Advent of Research Evaluation Systems. Dordrecht: Springer, 127-151.

Hirschauer, Stefan, 2002: Expertise zum Thema „Die Innenwelt des Peer Review. Qualitätszuschreibung und informelle Wissenschaftskommunikation in Fachzeitschriften." Förderinitiative »Wissen für Entscheidungsprozesse – Forschung zum Verhältnis von Wissenschaft, Politik und Gesellschaft« des BMBF. Online: http://www.sciencepolicystudies.de/dok/expertise-hirschauer.pdf.

Kieser, Alfred, 2010: Unternehmen Wissenschaft? Leviathan. Berliner Zeitschrift für Sozialwissenschaft 38, 347-367.

Lamont, Michèle, 2009: How Professors Think: Inside the Curious World of Academic Judgment. Cambridge: Harvard University Press.

Lamont, Michèle / Mallard, Grégoire / Guetzkow, Joshua, 2009a: Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review. Science, Technology & Human Values 34, 573-606.

Langfeldt, Liv, 2001: The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. Social Studies of Science 31, 820-841.

Leisyté, Liudvika / Boer, Harry de / Enders, Jürgen, 2010: Mediating Problem Choice: Academic Researchers' Responses to Changes in their Institutional Environment, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.):* a.a.O, 266-290.

Martin, Ben / Whitley, Richard, 2010: The UK Research Assessment Exercise. A Case of Regulatory Capture?, in: *Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.)*, a.a.O, 51-80.

Michaels, Axel, 2010: Rituale der Forschungsevaluation: Die große Begehung der Mittelbaustelle. Frankfurter Allgemeine Zeitung, 15. August 2010.

Neidhardt, Friedhelm, 2010: Selbststeuerung der Wissenschaft: Peer Review, in: *Simon, Dagmar / Knie, Andreas / Hornbostel, Stefan (Hg.):* Handbuch Wissenschaftspolitik. Wiesbaden: VS Verlag, 280-292.

Schimank, Uwe, 1995: Für eine Erneuerung der institutionalistischen Wissenschaftssoziologie. Zeitschrift für Soziologie 24, 42-57.

Schimank, Uwe, 2005: Die akademische Profession und die Universitäten: "New Public Management" und eine drohende Entprofessionalisierung, in: *Thomas Klatetzki /Veronika Tacke (Hg.):* Organisation und Profession. Wiesbaden: VS-Verlag, 143-164.

Torka, Marc, 2009: Die Projektförmigkeit der Forschung. Baden-Baden: Nomos.

Travis, G.D.L. / Collins, H.M., 1991: New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. Science, Technology, & Human Values 16, 322-341.

Weingart, Peter, 2005: Das Ritual der Evaluierung und die Verführbarkeit der Zahlen, in: *Ders. (Hg.):* Die Wissenschaft der Öffentlichkeit: Essays zum Verhältnis von Wissenschaft, Medien und Öffentlichkeit. Weilerswist: Velbrück, 102-122.

Whitley, Richard / Gläser, Jochen / Engwall, Lars (eds.), 2010: Reconfiguring Knowledge Production. Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford: Oxford University Press.