

---

KLAUS FISCHER

## Soziale und kognitive Aspekte des Peer Review-Verfahrens

Mit dem Peer Review System verhält es sich ähnlich wie mit dem Wetterbericht. In beiden Fällen handelt es sich um Komponenten komplexer rückgekoppelter Systeme, die sehr sensibel auf Eingriffe reagieren können. Zwar schert sich das Wetter im allgemeinen nicht um meteorologische Prognosen, während Wissenschaftler den *reviews* ihrer *peers* zuweilen mit deutlichen Emotionen begegnen, aber dennoch gehört die Wetterprognose im Zeitalter des „Klimaschutzes“ zum Systemzusammenhang, der das Wetter bestimmt. Wetterberichte können Menschen, Unternehmen und Staaten zu Handlungen veranlassen, die wiederum das zukünftige Wetter beeinflussen. Aneinandergereiht und aggregiert zu „Klimadaten“, bestimmen sie nicht nur die Agenda von Bürgerinitiativen, internationalen Organisationen und Weltkonferenzen, sie führen auch zu Eingriffen in ökologische Systeme, die das Wetter der Zukunft und a fortiori auch die zukünftigen Wetterberichte, aus denen das Kunstprodukt „Klima“ errechnet wird, verändern. Freilich besteht selbst für den extrem unwahrscheinlichen Fall, dass eines der aktuellen Klimamodelle stimmen sollte, keine Gewähr, dass die *erwünschten* Ergebnisse eintreten. Das Wetter ist ein chaotisches System, in dem in „kritischen Bereichen“ kleinste Schwankungen große Wirkungen ausüben können. Falsch berechnete Eingriffe ins „Klima“ können verheerende Folgen haben – nicht nur für das zukünftige Wetter, sondern vor allem für die Ökonomien der Länder, die die Eingriffe vornehmen.

Auch die Produkte des Peer Review Systems können als Berichte und Urteile über lokale Eigenschaften eines komplexen Systems aufgefasst werden – Berichte, die, indem sie Menschen zu bestimmten Handlungen veranlassen, dieses System auf allen Ebenen und Skalen seiner Organisation und Dynamik über informationelle, ökonomische, kulturelle oder politische Rückkopplungen verändern zu können – und die selbst wiederum von diesen Veränderungen beeinflusst werden. Nicht nur Artikel, Bücher und Projektanträge einzelner Wissenschaftler, auch Spezialgebiete, Disziplinen, Laboratorien, Fächer, Universitäten, Forschungverbände, Organisationen der Forschungsförderung und nationale Wissenschaftssysteme werden mittlerweile dem Urteil des Peer Review Systems unterworfen. Wie im Falle des Wetters und seines Derivats, des Klimas, können Fehleinschätzungen

und Messfehler verheerende Konsequenzen haben. Niemand sollte sich darauf verlassen, dass diese Folgen erst langfristig sichtbar werden. Da auch das Wissenschaftssystem chaotische Eigenschaften hat, können fehlerhafte Bewertungen – vorgenommen in kritischen Bereichen zur passenden Zeit und am passenden Ort – Kaskaden ungeplanter und nicht rücknehmbarer Schäden auslösen.

Jeder weiß, dass man sich auf den Wetterbericht nicht verlassen sollte, und nur die professionspolitisch motivierten Optimisten, die auf den Superrechner der neuesten Generation schielen, glauben noch, dass man aus der Aggregation von Wetterdaten ein chaotisches System wie das Klima berechnen kann. In einer solchen Situation kann die rationalste Handlungsoption darin bestehen, zwar weiter zu forschen, aber im übrigen das System möglichst wenig zu stören.

Von Seiten der Politik wird in dieser Situation in der Regel argumentiert, dass man auch bei unsicherer Beweislage handeln müsse, um Schaden abzuwenden. Das Peer Review System, so lautet der analoge Schluss, mag unvollkommen sein, aber es ist das einzig verfügbare Mittel, um den anschwellenden Informationsstrom zu filtern, die knappen Forschungsmittel zu kanalisieren und Defizite an unseren Wissenschafts- und Forschungseinrichtungen zu erkennen.

Wer diesem Argument zustimmt, sollte zumindest wissen, auf welches Risiko er sich einlässt und welchen Schaden er anrichten kann, wenn er der Funktionsweise dieses Systems blind vertraut. Wir werden im folgenden versuchen, die Zuverlässigkeit des Bewertungssystems der Wissenschaft im Lichte vorliegender Forschungsergebnisse zu prüfen und die Risiken von Fehlentscheidungen aufgrund wissenschaftsgeschichtlicher Erfahrungen und wissenschaftssoziologischer Überlegungen abzuschätzen.

Bevor wir das tun, müssen wir noch die Grundfrage behandeln, ob wir das Peer Review System überhaupt brauchen. Zumeist wird dies ohne nähere Prüfung vorausgesetzt. Doch welcher wissenschaftliche Durchbruch, welches jener intellektuellen Großereignisse, die den Lauf der Forschung tiefgreifender verändern als Tausende gestanzter und lackierter „Normprojekte“, wurde je vom Peer Review System erkannt und im Prozess seiner Formung *nicht* behindert – aus Missgunst, Hochmut, Ignoranz, dogmatischer Einstellung oder schlichter Inkompetenz? Wenn wir noch einmal unsere Analogie bemühen dürfen, so könnten wir sagen, dass es sich bei der Wettervorhersage oft nicht anders verhält, wenngleich die Ursachen des Versagens in der Regel andere sind. Die für viele Menschen einschneidendsten Wetterereignisse – der europaweite Orkan Ende der neunziger Jahre, das Hochwasser von 2002 in Sachsen und Thüringen, die europaweite Hitzewelle von 2003, katastrophale lokale Wetterereignisse (Tornados, Starkregen, extremer Hagel und Blitzschlag, etc.) – wurden von den Diensten nicht rechtzeitig vorhergesagt, das heißt, nicht früh genug, um das Handeln

danach ausrichten zu können. Wie wichtig ist ein Wetterbericht, der gerade die größten Wetterrisiken nicht vorhersagt? Wie wichtig ein Peer Review System, das gerade bei den wissenschaftlichen Großtaten kläglich scheitert?

Die beste Wetterprognose – so haben Statistiker einmal berechnet – ist die, die besagt, dass sich das Wetter nicht ändert, dass also das heutige Wetter auch morgen anhält. Übertragen auf wissenschaftliche Publikationen oder Forschungsvorhaben würde das bedeuten, dass es vermutlich keine schlechte Ausgangshypothese wäre, wenn wir davon ausgehen, dass dieselben Autoren oder Forscher im allgemeinen die gleiche Qualität liefern: Wer bisher gute Artikel schrieb oder gute Forschungsprojekte betrieb, wird dies vermutlich auch in Zukunft tun, und umgekehrt.<sup>1</sup>

Leider ist hieraus keine generelle Empfehlung zur Vereinfachung des Bewertungsverfahrens zu gewinnen, denn bereits die Prämisse des Arguments ist bestrittbar. Während Meteorologen relativ klare Maßstäbe zur Beschreibung des

1 Vgl. Abrams, P.A., The predictive ability of peer review of grant proposals: The case of ecology and the US National Science Foundation. – In: *Social Studies of Science*. 21 (1991), S. 111 – 132. Abrams glaubt, dass die Politik der NSF jeden Projektvorschlag für sich allein stehend, ohne Berücksichtigung der Vorgeschichte der Antragsteller, insbesondere ihrer bisherigen Forschungserfolge, zu bewerten, zu einer suboptimalen Verteilung der Forschungsgelder führt. Er weist nach, dass Forscher, die bisher erfolgreich waren, dies aller Wahrscheinlichkeit nach auch in der Zukunft sein werden und schlägt vor, diese Annahme zum Leitprinzip der Verteilung zu machen. Da sein Kriterium für Erfolg in Publikations- und Zitationsmaßen liegt, besteht dabei allerdings die Gefahr, dass innovative Forschung, die gegen den aktuellen Konsens betrieben wird und – wenn überhaupt – erst langfristig erfolgreich sein wird (und infolgedessen auch noch nicht mit hohen Zitationsziffern aufwarten kann), dabei benachteiligt wird. Richtig ist dagegen, dass die Bewertung der Projekte – und nicht der Antragsteller – auch dazu dienen kann, bestimmte Forschungsthemen, Forschungsprogramme und Forschungsmethoden nach politischen und nicht nach wissenschaftlichen Vorgaben zu fördern. Dies führt dazu, dass weniger kreative Wissenschaftler mit Erfolg Gelder für stromlinienförmige Projekte beantragen, während Wissenschaftler an der „cutting edge“ der Forschung Probleme haben, ihre Ideen zu verkaufen, weil sie zum einen schlecht in das vorgegebene thematische Raster passen und zum anderen, weil die Antragsteller oft nicht sagen können, was bei ihren Untersuchungen herauskommen wird (vgl. Donald Forsdyke, *Demographic shift in the scientific community*, <http://www.post.queensu.ca-forsdyke/peerrev.htm>). Ein damit zusammenhängendes Ergebnis erhalten Marsh, H.W. / Bazeley, P., *Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure*. – In: *Multivariate Behavioral Research*. 34(1999), S. 1 – 30. Die Autoren der auf Australien bezogenen Studie zeigen unter anderem, dass die Bewertungen von Forschergruppen und der von ihnen eingereichten Projektanträge hoch korreliert sind ( $r = .85$ ). Dies widerspricht den Argumenten von Abrams nur vordergründig, denn es waren die gleichen Gutachter, die die Projektanträge und die damit verbundenen Forschergruppen einschätzen mussten. Es handelte sich also nicht um unabhängige Qualitätsmaße.

Wetters von gestern haben, können Wissenschaftler sich oft nicht darüber einig sein, ob die bisherigen Arbeiten von A alle gut, die von B immer nur mittelmäßig und die von C durchgängig schlecht waren. Doch selbst wenn es so wäre – jeder Wissenschaftler fängt irgendwann an zu forschen und zu publizieren. Zumindes für die erste Arbeit müsste das Bewertungsproblem also auf andere Weise gelöst werden. Zweitens muss jeder die Chance haben, sich zu verbessern oder zu verschlechtern. Forschung hat auch eine handwerkliche Seite und in vielen Arbeiten spielt diese sogar eine entscheidende Rolle. Und schließlich strömen nicht jedem Forscher gleichermaßen publikations- oder förderungswürdige Ideen in kontinuierlicher Folge zu, obwohl die Zwänge der Forschungsfinanzierung und die Erwartungen der Peers Stetigkeit einfordern. Unter dem Diktat knapper Mittel und unter Berücksichtigung des daraus folgenden Allokationsproblems hat es also seinen guten Sinn zu fordern, Forschungsgelder und Publikationsraum erst nach sorgfältiger qualitativer Bewertung zu vergeben.

Den erhofften Vorteilen eines Systems der Forschungsbewertung stehen schwer kalkulierbare Kosten gegenüber. Aus der Wissenschaftsgeschichte, auch der jüngsten, ist bekannt, dass gerade die sehr innovativen Denker, die außerhalb des *mainstream* forschen und von deren Arbeit die großen qualitativen Fortschritte der Wissenschaft abhängen, zumindest zeitweilig auf besondere Schwierigkeiten der Anerkennung gestoßen sind. Kopernikus, Bruno, Galilei, Semmelweis, Robert Mayer, Gregor Mendel – beinahe jedes Kind kennt die Symbolgestalten der bisherigen Kämpfe der wissenschaftlichen Vernunft gegen die vernagelten Vertreter der Tradition, gegen die Blockierer, die den Fortschritt aufhalten wollten und ihn in ihrer Borniertheit nicht einmal erkannten. Natürlich ist dieses Bild eine romantische Idealisierung, die allerdings insofern einen wahren Kern hat, als sich das Neue oft nur nach schweren Konflikten durchsetzt, in deren Verlauf seine Fürsprecher für das Überschreiten geistiger Grenzen, die zugleich den geistigen Horizont ihrer Kritiker markieren, schwer büßen müssen.

Wie die „wissenschaftliche Gemeinschaft“, so ist auch das Peer Review System als dessen Teil bei der gerechten Bewertung von Neuerern eines gewissen Formats offenbar überfordert. Doch wie schneidet es im Normalfall ab, also bei der Beurteilung üblicher und durchschnittlicher Leistungen, die quantitativ den Löwenanteil aller Fälle ausmachen? Lehrer und Gutachter müssen bewerten, dies ist eine der Aufgaben, für die sie bestellt sind. Täglich werden an Schulen und Universitäten Tausende von Aufsätzen, Diktaten, Referaten, Klausuren, Hausarbeiten und Abschlussarbeiten benotet. Sind diese Bewertungen objektiv und reliabel?

Empirische Untersuchungen zeigen, dass auch die Bewertungen von Schüler- und Studentearbeiten nur begrenzt reproduzierbar sind. Je mehr es der Gutach-

ter mit Leistungen zu tun hat, die an die Grenze des Wissens und der Forschung vorstoßen, also mit Dissertationen, Habilitationsschriften, Forschungsberichten von Kollegen („peers“), desto schwieriger wird seine Aufgabe und desto riskanter seine Bewertung. Es kann daher nicht überraschen, dass diese Bewertungen, technisch gesprochen, nur eine bescheidene Reliabilität (Übereinstimmung zwischen verschiedenen Bewertern = Inter-Bewerter-Übereinstimmung) aufweisen. Einzelne Ergebnisse deuten sogar darauf hin, dass nicht einmal Objektivität (Intra-Bewerter-Übereinstimmung) gegeben ist. Zu verschiedenen Zeiten, in unterschiedlichen Stimmungen, bei anderer zeitlich vorangehender Lektüre – also bei Verschiebung der Bezugsebene – kann sich das Urteil eines Gutachters über den gleichen Aufsatz oder den gleichen Projektantrag sehr verschieden darstellen.<sup>2</sup>

Doch wie steht es um die dritte, die wichtigste Dimension einer Messung, ihre Validität? Inwiefern Objektivität und Reliabilität im genannten Sinne notwendig sind, um die Validität des Begutachtungsprozesses zu sichern, ist eine kaum eindeutig zu beantwortende Frage. Validität weist ein Begutachtungssystem dann auf, wenn die von ihm getroffenen Entscheidungen das befördern, was den Sinn von Forschung ausmacht, nämlich den Fortschritt unserer Erkenntnis. Eine hohe Reliabilität wird diesem Ziel genau dann *nicht* dienen, wenn die Grundlage der Einmütigkeit in gemeinsam geteilten Vorurteilen<sup>3</sup> (etwa denen einer paradigmageleiteten Gruppe), in gemeinsamen sozialen, ökonomischen oder politischen Interessen, in einer gemeinsamen Ideologie oder Weltanschauung besteht. Ob dies der Fall ist, kann der Wissenschaftshistoriker oft nur im Rückblick entscheiden. Die Forderung nach hoher Reliabilität (und als Voraussetzung für diese, nach hoher Objektivität) ist daher zweischneidig. Reproduzierbarkeit und Verlässlichkeit des Bewertungsprozesses sind somit nur Aspekte einer komplexeren Problematik.

Das Peer Review System hat sowohl für die Verteilung von Forschungsmitteln als auch für die Publikation der Ergebnisse eine Schlüsselfunktion.<sup>4</sup> Es ist der Filter, der darüber entscheidet, welche Projekte es verdienen, gefördert zu werden, und welche Informationen es wert sind, von der „wissenschaftlichen Gemein-

2 Warum dies so ist, wird erklärt in: Laming, D., Why is the reliability of peer review so low? – In: *The Behavioral and Brain Sciences*. 14(1991), S. 154 – 156. Aus unterschiedlichen Experimenten zu Wahrnehmung und Urteilsbildung scheint sich zu ergeben, dass es einen Basiseffekt gibt, der etwa zwei Drittel der Varianz bestimmt. Dieser Basiseffekt wird durch die Verschiebung des Referenzsystems erzeugt und hat nichts mit den objektiven Eigenschaften des zu bewertenden Items zu tun. In dieselbe Kategorie fällt die Beobachtung, dass es einen großen Unterschied für die Erfolgsaussichten bedeuten kann, ob ein Projektantrag zu Beginn der Sitzung der Vergabe-Kommission oder eher am Ende behandelt wird.

3 Beispiele in: Schönemann, P.H., In praise of randomness. – In: *The Behavioral and Brain Sciences*. 14(1991), S. 162 – 163.

schaft“ zur Kenntnis genommen zu werden. Letztthin entscheidet das Peer Review System über Schicksale und Lebensläufe, über Karrieren, Reputation, Ruhm und Auszeichnungen im Wissenschaftssystem. Wie in einigen anderen Ländern wird das Peer Review System zukünftig auch in Deutschland über Leistung und Gehalt, über die Finanzierung von Arbeitsgruppen, Instituten, Fächern und Universitäten entscheiden. An ein Instrument, das derart einschneidende Entscheidungen begründen soll, sind sehr hohe Anforderungen zu stellen. Anders formuliert, das Bewertungssystem der Wissenschaft braucht härtere Kriterien als nur das persönliche Gefühl, die subjektive Relevanzhierarchie oder den siebten Sinn des Urteilenden. Die entscheidende Frage lautet:

1. *Verfügt das Peer Review System über zuverlässige Kriterien der Bewertung?*

Selbst die Herausgeber wissenschaftlicher Zeitschriften reagieren auf diese Frage skeptisch. Drummond Rennie, Deputy Editor der Zeitschrift JAMA, des Organs der American Medical Association, äußerte aus Anlass des „Fourth International Congress on Peer Review in Biomedical Publication“, beim Peer Review handle es sich um „a system we know to be time-consuming, complex, expensive and [...] prone to abuse, while we acknowledge that the scientific evidence for its value is meager. Indeed, if the entire peer-review system did not exist but were now to be proposed as a new invention, it would be hard to convince editors looking at the evidence to go through the trouble and expense.“<sup>5</sup>

Wie sieht die angesprochene Evidenz aus?

Da systematische Studien zur Zuverlässigkeit des Peer Review Systems aus nachvollziehbaren Gründen auf große Schwierigkeiten stoßen, gibt es bis heute kein statistisch zuverlässiges Gesamtbild über die Güte dieses Verfahrens. Die vorliegenden sporadischen Untersuchungen ersetzen keine Gesamtdarstellung, aber sie erlauben dennoch Folgerungen, die als alarmierend betrachtet werden müssen, obwohl sie möglicherweise noch ein zu positives Bild der Gesamtheit zeigen. Die Untersuchungen waren von der Zustimmung von Zeitschriften oder Drittmittelgebern abhängig, und diese war um so leichter zu erwarten, je mehr

4 „Clearly, peer review is the ‘sacred cow’ of S & T metrics – the traditional ‘right’ of scientists to critique and review the work of their colleagues. It is also at the core of the method by which science progresses.“ (Geisler, E., *The Metrics of Science and Technology*. Westport, Connecticut & London: Quorum Books 2000. S. 234).

5 Rennie, D., *Fourth International Congress on Peer Review in Biomedical Publication* (Editorial). – In: *Journal of the American Medical Association*. 287(2002)21, S. 2760.

die Entscheidenden von der Güte ihres Systems überzeugt waren. Ohne Wissen der Zeitschriften unternommene Feldexperimente, Erfahrungsberichte und theoretische Überlegungen runden das Bild ab. Die Ergebnisse widersprechen – mit wenigen Ausnahmen – dem Bild einer Wissenschaft, die ihre Qualitätsmaßstäbe nicht nur kennt, sondern sie auch umzusetzen weiß, die nicht nur die soliden, sondern auch die zukunftsweisenden Arbeiten erkennt und die redundanten oder mangelbehafteten vor ihrer Publikation ausscheidet.

Den Wissenschaftshistoriker, der sich mit Wissenschaft als sich entwickelndes System beschäftigt, konnte das nicht überraschen. Aus Biographien und Autobiographien von Gelehrten, Forschern und Erfindern kennen wir viele Beispiele, in denen das Peer Review System krasse Fehlurteile über Forschungsspitzenleistungen gefällt hat. In einer Reihe von Fällen haben Gutachter von Fachzeitschriften zentrale Arbeiten späterer Nobelpreisträger abgelehnt, weil sie deren Qualität nicht erkannt haben. Beispiele sind Enrico Fermi, Sir Hans Krebs, Rosalyn Yalow<sup>6</sup>, Gerd Binnig und Hans Rohrer (die Erfinder des Rastertunnel-Mikroskops)<sup>7</sup>, die Erstentdecker des Top Quarks<sup>8</sup> und viele andere. Bekannte Wissenschaftler, deren Ideen zunächst auf Ignoranz oder vehemente Ablehnung seitens der „wissenschaftlichen Gemeinschaft“ stießen, sind Alfred Wegener, Alan Turing, Konrad Zuse, Hermann Oberth, Peyton Rous, Mitchell Feigenbaum, Frank Rosenblatt, Stanley Prusiner, Günter Blobel, Andrei Linde, Noam Chomsky, Karl Popper.<sup>9</sup> Es trifft in erster Linie jüngere, noch unbekannte, nichtetablierte Wissenschaftler, die sich zu weit vom aktuellen Konsens entfernt haben, ferner solche, die die Grenzen ihrer Disziplin verletzen. Hin und wieder trifft es auch etablierte Wissenschaftler, selbst Nobelpreisträger, wenn sie sich zu weit vom Konsens der Mehrheit entfernen (z.B. Brian Josephson), oder wenn sie zu offen

- 6 Yalow, R., Competency testing for reviewers and editors. – In: *The Behavioral and Brain Sciences*. 5 (1982), S. 244; Cicchetti, D.V., Referees, editors, publication practices. – In: *Science and engineering Ethics*. 3(1997), S. 58. Cicchetti nennt noch ein anderes Beispiel: „The manuscript presenting the discovery of blood-typing was rejected by the highly regarded journal *Lancet*.“ (a.a.O.).
- 7 Armstrong, J.S. / Hubbard, R., Does the need for agreement among reviewers inhibit the publication of controversial findings. – In: *The Behavioral and Brain Sciences*. 14(1991), S. 136. Das Manuskript mit den ersten vom Rastertunnel-Mikroskop erhaltenen Resultaten wurde abgelehnt, weil ein Gutachter das Manuskript „not interesting enough“ fand (a.a.O.).
- 8 Graßmann, H., *Das Top Quark, Picasso und Mercedes Benz, oder: Was ist Physik?* Berlin: Rowohlt 1997. S. 178 f.
- 9 Weitere Informationen und Literatur dazu in: Fischer, K., *Ist Evaluation unvermeidlich innovationshemmend?* – In: Eveline Pipp (Hg.), *Drehscheibe E-Mitteuropa. Information: Produzenten, Vermittler, Nutzer. Die gemeinsame Zukunft (Biblos-Schriften Band 173)*. Wien: Phoibos 2002. S. 109 – 128.

zu verstehen geben, dass sie selbst nicht wissen, was bei ihren Untersuchungen herauskommen wird.<sup>10</sup> Aber die Extremfälle markieren vielleicht nur die Spitze des Eisberges. Welches Vertrauen – so muss man fragen – verdient ein System, das nicht einmal die absolute Qualitätsspitze verlässlich erkennt?

Unterdessen gibt es Feldexperimente, die den Versuch machen, die Gründe für dieses Versagen des wissenschaftlichen Bewertungsmechanismus zu bestimmen. Das klarste und zugleich vernichtendste ist das von Peters und Ceci 1982 beschriebene.<sup>11</sup> Peters und Ceci reichten bei zwölf englischsprachigen und hoch angesehenen psychologischen Fachzeitschriften je einen Artikel zur Publikation ein. Was sie den Herausgebern der jeweiligen Zeitschrift nicht sagten, war, dass genau dieser Artikel in der gleichen Zeitschrift bereits 18 bis 32 Monate vorher schon einmal erschienen war. Die ausgewählten Artikel waren von Mitgliedern der Top-Departments des Faches an amerikanischen Universitäten verfasst. Um die Aufgabe der Herausgeber und Gutachter zu erschweren, hatten Peters und Ceci allerdings die Namen und die Arbeitsstätten der Autoren geändert und auch die Zusammenfassung mit den Schlüsselwörtern modifiziert. Das Ergebnis war erstaunlich: Nur drei der zwölf eingereichten Artikel wurden in einen oder anderen Stadium des Bewertungsprozesses als bereits erschienen erkannt. Von den restlichen neun wurden acht abgelehnt, zum Teil aufgrund von „serious methodological flaws“. Ein einziger Artikel wurde zur Publikation angenommen. Da die durchschnittliche Ablehnungsquote bei den betreffenden Zeitschriften etwa 80% betrug, können wir schließen, dass die sich ergebende Verteilung im Rahmen der Erwartungen lag – wäre da nicht die störende Tatsache, dass alle Artikel in einem vorangehenden Begutachtungsverfahren von den gleichen Zeitschriften bereits akzeptiert worden waren. Offenbar hatte das Bewertungssystem in diesem Fall keine inhärenten und auch nur im bescheidensten Sinne objektiven Maßstäbe, um ein reproduzierbares Urteil fällen zu können.<sup>12</sup>

An den Artikel von Peters und Ceci schloss sich eine lange und scharfe Auseinandersetzung an. Eine der vorgebrachten Erklärungen für das Ergebnis des Versuchs beleuchtet vor allem die soziale Komponente des Bewertungsprozesses.

- 10 Natürlich ist gerade diese Ergebnisoffenheit das Kennzeichen innovativer Forschung, aber sie verursacht Unbehagen bei jenen, die immer alles „unter Kontrolle“ haben möchten, weil sie – zum Beispiel – anderen gegenüber rechenschaftspflichtig sind. Diese Struktur der Forschungsförderung trägt der Funktionsweise erfolgreicher Wissenschaft, die immer chaotische – also nichtberechenbare – Eigenschaften aufweisen muss, nicht angemessen Rechnung.
- 11 Peters, D.P. / Ceci, S.J., Peer-review practices of psychological journals: The fate of published articles, submitted again. – In: *The Behavioral and Brain Sciences*. 5(192), S. 187 – 195. Der Aufsatz war vorher von den Zeitschriften *Science* und *American Psychologist* abgelehnt worden (vgl. Armstrong, J.S., Barriers to scientific communication: The author's formula. – In: *The Behavioral and Brain Sciences*. 5(1982), S. 197).

Ein wichtiger Teil des Feldexperiments hatte darin bestanden, prestigeträchtige Namen von Autoren und Institutionen durch frei erfundene und deswegen unbekannte zu ersetzen. Einige Kommentatoren kritisierten diese Versuchsbedingung. Sie hielten es für legitim, dass Gutachter sekundäre Kriterien wie den Bekanntheitsgrad der Autoren oder das Ansehen der Heimatinstitution der Autoren als Indikatoren für wissenschaftliche Qualität benutzten. Dem liegt die Vermutung zugrunde, dass hinter dem sozialen Merkmal wieder kognitive Qualitäten stehen, die man nicht direkt – zum Beispiel an einem Manuskript – wahrnehmen kann. Fehlende kognitive Kriterien werden durch soziale substituiert. Dies ist eine Form der Bewertung, die nach dem Muster verfährt: X ist eine Spitzenuniversität, Spitzenuniversitäten stellen nur hervorragende Leute ein, hervorragende Leute schreiben gute Artikel. Andererseits: Wer kennt schon das „Tri-Valley Center for Human Potential“? Kann jemand gut sein, der an einer Institution dieses Namens arbeitet?<sup>13</sup> Kritiker dieser Art von Analogieschluss argumentieren, dass ein solcher Zusammenhang, wenn er denn bestehen sollte, bestenfalls statistischer Natur sei und keine Rückschlüsse auf den Einzelfall zulasse. Das ernüchternde Resultat des von Peters und Ceci durchgeführten Feldexperiments lautet, dass dann, wenn dieses sekundäre Indiz für Qualität entfällt, die psychologischen Gutachter blind für die inhärenten kognitiven Qualitäten einer Arbeit zu werden scheinen.<sup>14</sup> Die Reaktion von John Ziman zeigt die Bestürzung, aber auch die Ratlosigkeit vieler Kommentatoren über dieses Resultat an. „The consensus of reviewers *against* the resubmitted papers suggests something worse than total chaos – but that is bad enough to make nonsense of the whole system. The peer-review

- 12 Ähnliche Experimente sind aus der Literatur bekannt. Chuck Ross beschreibt eines, das er selbst unternommen hat. „In 1977 I did an experiment similar to Peters & Ceci’s (...) but in the world of mainstream fiction. I typed up Jerzy Kosinskys (1968) *Steps* and submitted it, untitled, to 14 major publishing houses and 13 literary agents. To another 13 agents I sent a letter of inquiry. The highly acclaimed novel, which had won the prestigious National Book Award for fiction in 1969, was rejected by all (including Random House, its original publisher). No one recognized the work, and no one thought it deserved to be published.“ (Ross, C., Rejecting published work: Similar fate for fiction. – In: *The Behavioral and Brain Sciences*. 5(1982), S. 236.) Wenn dieses Ergebnis verallgemeinerbar sein sollte, dann würde dies bedeuten, dass literarischer Ruhm ebenso unverdient ist wie literarisches Scheitern. Dass die Selektion wissenschaftlicher Artikel zumindest im oberen Qualitätssegment (was immer das bedeuten mag) starke Ähnlichkeiten mit Urteilen über Kunstwerke hat, betont der Soziologe Duncan Lindsey. (Precision in the manuscript review process: Hargens and Herting revisited. – In: *Scientometrics*. 22(1991), S. 314.) Wenn Lindsey recht hat, dann wäre in der Tat vom Peer Review System kein höheres Maß an Objektivität, Reliabilität und Validität zu erwarten als das bei der ursprünglichen Rezeption von Kunstwerken (Musik, Dichtung, Bildende Kunst) zu beobachtende. Siehe dazu: Roh, E., *Der verkannte Künstler. Studien zur Geschichte und Theorie des kulturellen Missverstehens*. Köln: DuMont 1993 (orig. 1938).

process seems not merely imperfect: It is an entirely useless, if not positively harmful activity, based upon quite erroneous assumptions. I recoil from this conclusion, not because it is inconceivable but because it would take a very long time to imagine what to say or do next.”<sup>15</sup>

Andere Untersuchungen mit unterschiedlichem Design haben eine bescheidene bis mittlere Übereinstimmung zwischen Erstbegutachtung und Replikation, bzw. zwischen den Beurteilungen durch verschiedene Gutachter erbracht. Dieses Resultat, das auf eine partielle, aber dennoch relativ geringe Objektivität und Reliabilität verweist, ist allerdings noch kein hinreichender Beleg für die Validität des Verfahrens, sondern kann auch durch das zeitweise Vorherrschen gewisser konsensueller Sichtweisen erklärt werden. Wir kommen darauf noch zurück.

Dass die Bewertungsstandards in den Geisteswissenschaften zumindest nicht höher als in der Psychologie sind, wurde in einem Feldversuch bestätigt, der dem Experiment von Peters und Ceci nachempfunden ist. Der – fiktive – Wiener Realschulprofessor Eduard Kindler schickte im Februar 1996 zehn deutschsprachigen Philosophiezeitschriften einen Aufsatz mit dem Titel „Wissenschaftliche Probleme der Naturphilosophie“ mit der Bitte um Publikation. Was die angeschriebenen Zeitschriften nicht wussten, war, dass der Text des Aufsatzes identisch war mit einem sehr bekannten und häufig nachgedruckten Artikel von Karl Popper, dessen deutsche Fassung den Titel „Die Zielsetzung der Erfahrungswissenschaft“ trägt. Das nun nicht mehr sonderlich überraschende Ergebnis war, dass neun von zehn Zeitschriften den Artikel ablehnten – aus qualitativen Gründen und nicht etwa, weil sie den Betrug erkannt hatten.<sup>16</sup>

- 13 Auch in der Physik ist dieses Denkmuster bekannt, wie an der Stellungnahme von Robert K. Adair, Herausgeber des vielleicht prestigeträchtigsten Organs des Fachs, *Physical Review Letters*, zu der Untersuchung von Peters und Ceci zu ersehen ist. „Science is not democratic, and it is neither unnatural nor wrong that the work of scientists who have achieved eminence through a long record of important and successful research is accepted with fewer reservations than the work of less eminent scientists”. Adair, R.K., A physics editor comments on Peters and Ceci’s peer-review study. – In: *The Behavioral and Brain Sciences*. 5(1982), S. 196. Die Nobelpreisträgerin Rosalyn Yalow stützt diese Argumentation so: „I am in full sympathy with rejecting papers from unknown authors working in unknown institutions. How does one know that the data are not fabricated? Those of us who publish establish some kind of track record. If our papers stand the test of time and are shown to be valid through confirmation by other investigators, it can be expected that we have acquired expertise in scientific methodology. Admittedly this is not always so.” (Yalow, R., Competency testing for reviewers and editors. – In: *The Behavioral and Brain Sciences*. 5(1982), S. 244). Doch wenn das Argument nur statistisch gilt, wie die Autorin im weiteren Verlauf ihrer Ausführungen zugesteht, wäre es ungerecht und methodisch fehlerhaft, die Arbeiten unbekannter Forscher von unbekanntem Institutionen pauschal abzulehnen. Für eine solche Strategie braucht man kein Peer Review System. Seine Funktion besteht gerade darin, die *inhärenten Qualitäten* einer Arbeit oder eines Projektantrages zu bewerten.

Dieses Experiment ist insofern bemerkenswert, als es nicht nur die Reliabilität, sondern auch die Validität des Begutachtungssystems betrifft. Schließlich hat sich der genannte Aufsatz innerhalb einer bestimmten wissenschaftsphilosophischen Tradition bewährt und wird von ihr als „Klassiker“ anerkannt. Als Teil einer sedimentierten Tradition ist der Essay allerdings in einem wohlverstandenen Sinne nicht auf dem „Stand der aktuellen Diskussion“. Dies kann zweierlei bedeuten. Es kann bedeuten, dass der Artikel von Fachleuten begutachtet wurde, die erkannt haben, dass er für den Spezialisten ungeachtet der Richtigkeit der vorgebrachten Argumente kaum Neues bringt. Es kann aber auch heißen, dass der Artikel einfach dem heutigen philosophischen Geschmack nicht mehr entspricht: Es haftet ihm der Geist einer anderen Zeit an, der philosophische Trend-

- 14 Kritiker, die selbst Herausgeber wissenschaftlicher Zeitschriften waren oder sind, stimmen der Forderung nach möglichst hoher Übereinstimmung zwischen den Gutachtern oft nicht zu. „As an editor, I intentionally act in ways that lower the correlation between ratings. For example, I give the manuscript to reviewers who have very different strengths or skills to bring to the manuscript.“ (Kiesler, C.A., Confusion between reviewer reliability and wise editorial and funding decisions. – In: *The Behavioral and Brain Sciences*. 14(1991), S. 151.) John C. Balair verweist darauf, dass Gutachten nur Entscheidungshilfen sind, die erst durch den Herausgeber einer Zeitschrift zu einem Gesamtbild integriert werden müssen, wobei je nach Breite einer Zeitschrift ein komplettes „Editorial Board“ notwendig sei. „There is no substitute for careful study of specific comments, integrated with the wisdom of editorial board members and, sometimes, special consultants. As a result, it was not unusual for us to publish papers that three reviewers had recommended for disapproval, and vice versa.“ (Balair, J.C. Reliability, fairness, objectivity and other inappropriate goals in peer review. – In: *The Behavioral and Brain Sciences*. 14(1991), S. 138.) Konsequenterweise muss man zu dem Schluss kommen, dass im Feldexperiment von Peters und Ceci nicht die Gutachter, sondern die Herausgeber versagt haben. Kritisch wäre zu dieser Interpretation anzumerken, dass sie ein idealisiertes Bild eines Herausgebers präsentiert, der gewissermaßen – aufgrund seiner langen Erfahrung – über den Dingen steht und sie deshalb besser beurteilen kann. Dies mag im Einzelfall durchaus so sein. Aber es wäre gefährlich, die Güte des Systems auf kontingente persönliche Qualitäten der Herausgeber zu stützen und strukturelle Bedingungen zu ignorieren. Grundsätzlich gilt, dass Herausgeber ungeachtet ihres punktuellen Informationsvorsprungs natürlich Teil des Bewertungssystems bleiben und somit den sozial-kognitiven Mechanismen der wissenschaftlichen Gemeinschaft unterliegen. Wenn sie von den Gutachtern abgelehnte Arbeiten dennoch publizieren, dann heißt dies nicht, dass sie die Bewertungen der Zukunft generell besser vorhersehen können als die Gutachter, sondern dass sie eine zeitschriftenspezifische *Fachinformationspolitik* betreiben. Damit soll nicht gesagt sein, dass die persönlichen Eigenschaften eines Herausgebers unwichtig sind. Der Autor glaubt im Gegenteil, dass die Fachkompetenz eines verantwortlichen Herausgebers und sein in langer Erfahrung erworbenes „Gespür“ für Qualität unverzichtbar für eine gute Editionsarbeit sind.
- 15 Ziman, J., Bias, incompetence, or bad management? – In: *The Behavioral and Brain Sciences*. 5(1982), S. 245.
- 16 Jung, J., *Der Niedergang der Vernunft. Kritik der deutschsprachigen Universitätsphilosophie*. Frankfurt: Campus 1997. S. 67ff.

denker kann ihn nicht als Teil des „aktuellen Diskurses“ einordnen.<sup>17</sup> Es ist wichtig, diese beiden Aspekte auseinanderzuhalten, weil sich aus ihnen konträre Folgerungen hinsichtlich der Validität des Begutachtungswesens deutscher philosophischer Fachzeitschriften ergeben.

Wieder dem Bereich der Psychologie entstammt ein anderes Feldexperiment, in dem gezeigt wurde, dass Gutachter dazu tendieren, identische Arbeiten positiver zu bewerten, wenn die darin erhobenen Daten mit ihren eigenen Ansichten übereinstimmen. Im Gegenzug tendieren sie zu einer kritischeren Einstellung zu solchen experimentellen Methoden, die Daten ergeben, die mit ihren eigenen theoretischen Einstellungen in Konflikt kommen.<sup>18</sup> Im gleichen Artikel wird nachgewiesen, dass auch die Nationalität und das Ansehen der Heimatuniversität des Autors eine signifikante Rolle bei der Bewertung wissenschaftlicher Leistungen (in diesem Fall im Bereich der Sozialwissenschaften) spielen.

Einer der Teilnehmer eines Symposiums über das Begutachtungssystem der drei großen amerikanischen Fachzeitschriften für Soziologie, ASR, AJS und Social Forces, kam zu dem Schluss „that the reviewers’ attachment to a given para-

Tabelle 1: *Nationalität und Ansehen der Heimatuniversität als Faktor von „reviewer bias“: Anteile eingereicherter Manuskripte, die als „gut“ bewertet wurden.*

	U. K. Authors	North American Authors	Totals
U. K. referees	70% of 600 papers	65% of 307 papers	907
North American referees	60% of 35 papers	75% of 20 papers	55
Totals	635	327	962
	Minor university authors	Major university authors	Totals
Minor university referees	65% of 120 papers	68% of 80 papers	200
Major university referees	55% of 110 papers	82% of 309 papers	419
Totals	230	389	619 <sup>a</sup>

a.Gordon, Refereeing reconsidered, op. cit., 234

- 17 Von den bei Jung auszugsweise abgedruckten Antworten begründet nur eine die negative Stellungnahme damit, dass der Artikel „für ein Fachjournal etwas zu allgemein und nicht auf dem gegenwärtigen Stand der Diskussion sei“. (Jung, op. cit., 70.)
- 18 Gordon, M., Refereeing reconsidered: An examination of unwitting bias in scientific evaluation. – In: Balaban, M. (ed.), Scientific Information Transfer: The Editor’s Role. Dordrecht und Boston: Reidel 1978. S. 231 – 235.

digm and the way one should proceed will strongly color their evaluation of a manuscript [...] The discipline does not provide the sociologist with a clear specification of what a good problem is, how the problem should be approached, what good evidence is, and what the proper techniques for analysis are”<sup>19</sup> ...

Eine aus gewöhnlich gut unterrichteten Kreisen kolportierte Geschichte untermauert die Bedeutung sekundärer Kriterien für den Begutachtungsprozess: Nachdem die *Zeitschrift für Soziologie* versuchsweise ein Dreifachblindverfahren<sup>20</sup> der Begutachtung eingeführt hatte, ergab sich der beunruhigende Befund, dass plötzlich auch Arbeiten soziologischer Meisterdenker durchfielen. Anstatt zu schließen, dass auch ein „Meisterdenker“ einmal eine unmeisterliche Arbeit abliefern konnte, modifizierte man das Verfahren so, dass Pannen dieser Art nicht mehr vorkommen konnten.

Die Bedeutung der institutionellen Anbindung von Autoren und Gutachtern wurde in einer Analyse der Begutachtungsprozesse dreier amerikanischer Zeitschriften aus dem Bereich der Sozialwissenschaften bestätigt. „It was then found that as the proportion of evaluators chosen from certain groups of institutions increased, so the proportion of successful authors from those groups of institutions similarly increased.”<sup>21</sup>

Peer Review in der Forschungsförderung und Peer Review im Publikationssystem der Wissenschaft sind ähnlich strukturiert und stehen deshalb ähnlichen Problemen gegenüber. Im Unterschied zur Begutachtung von Artikeln hat es das Peer Review System im ersten Fall mit Projektplänen zu tun, bei denen das Ergebnis noch nicht vorliegt. Insofern ist das Risiko möglicher Fehlentscheidungen noch höher als im zweiten Fall. Andererseits können negative Entscheidungen über Forschungsmittel für die Betroffenen weitaus größere existentielle Konsequenzen haben als gescheiterte oder verzögerte Publikationsvorhaben. Im Extremfall – bei großer Abhängigkeit von begutachteten Drittmittelprojekten – stehen Wissenschaftskarrieren, die Existenz von Forschungsgruppen und Instituten auf dem Spiel. Es ist deshalb wichtig zu wissen, wie das Peer Review System in der Forschungsförderung arbeitet und welche Konsequenzen aus möglichen Funktionsfehlern zu ziehen sind.

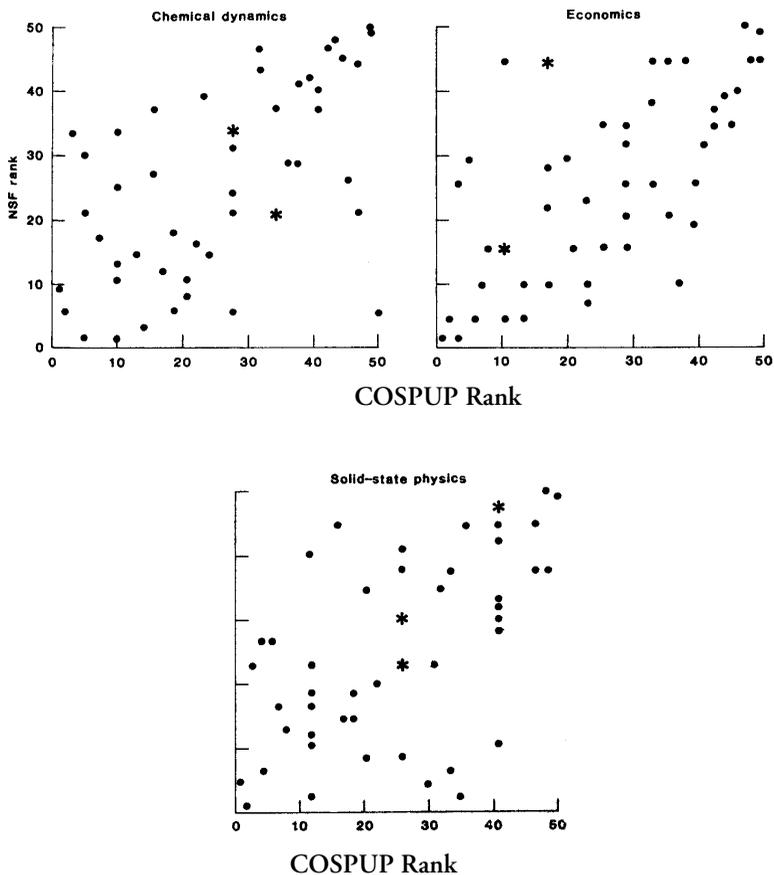
In einer vielbeachteten und heftig diskutierten Studie haben Stephen Cole, Jonathan Cole und Gary Simon 150 Projektanträge bei der amerikanischen National Science Foundation erneut begutachten lassen. Als Ergebnis erhielten sie,

19 Gove, W.R., The review process and its consequences in the major sociology journals. – In: *Contemporary Sociology*. 8(1979)6, S. 801.

20 Dabei kennen weder die Herausgeber noch die Gutachter die Namen der Autoren; die Autoren kennen, wie üblich, auch nicht die Namen der Gutachter.

21 Gordon, M., Refereeing reconsidered, op. cit. S. 232; siehe Tabelle 1.

Abbildung 1: *Streudiagramme in der Reliabilitätsstudie von Cole, Cole & Simon*  
*Rank order of proposals according to mean ratings NSF and COSPUP reviewers.*  
*N = 50 in each program. \* Asterisk indicates two proposals with identical ranks.*



“that getting a research grant depends to a significant extent on chance. The degree of disagreement within the population of eligible reviewers is such that whether or not a proposal is funded depends in a large proportion of cases upon which reviewers happen to be selected for.”<sup>22</sup> Die Streudiagramme für die jeweils 50 Anträge aus den Fächern Chemische Dynamik, Ökonomie und Festkörperphysik geben einen optischen Eindruck davon, wie sich die ursprünglichen Urteile von den Zweitbegutachtungen unterscheiden, vgl. Abbildung 1.

Tabelle 2: *„Subjektive“ Varianzanteile der ursprünglichen und replizierten Begutachtungen in der NSF Studie von Cole, Cole & Simon (Percentage of total variance in reviewers' ratings accounted for by difference among reviewers of individual proposals. The number in parentheses is the total number of reviewers. For each field there were 50 proposals.)*

Wissenschaftsgebiet	Prozent Varianz NSF	Prozent Varianz COSPUP
Chemische Dynamik	60 (242)	53 (213)
Ökonomie	51 (192)	49 (190)
Festkörperphysik	43 (163)	47 (182)

Wenn man es quantifizieren will, dann könnte man sagen, dass im Durchschnitt der untersuchten Fächer etwa 50% der Varianz durch Faktoren bestimmt werden, die sich – neutral ausgedrückt – als Dissens der Gutachter niederschlagen. Wäre der Gutachterprozess vollständig durch Zufall bestimmt, so würden sich bei jedem neuen Durchgang etwa 50% der Urteile umkehren. Die empirisch messbare „reversal rate“ beträgt demgegenüber etwa 25%. Daraus schließen sie, dass das Urteil der Gutachter zu etwa 50% auf „objektiven Gründen“ beruht. „Objektiv“ heißt dabei allerdings nur, dass über die angewandten Kriterien aktuell ein Konsens zu bestehen scheint.

Wir können hier die empirischen Resultate zu Funktionsweise und Leistung des Peer Review Verfahrens nur sehr selektiv darstellen.<sup>23</sup> Einige der herausragenden und offenbar „harten“ Ergebnisse sind die folgenden:

- Gutachter sind sich häufig uneinig, und zwar nicht nur in Einzelfragen, sondern auch in der grundlegenden Bewertung eines Manuskripts oder Projektplanes.<sup>24</sup>

22 Cole, St. / Cole, J.R. / Simon, G.A., Chance and consensus in peer review. – In: Science. 214 (1981), S. 881. Die folgende Abbildung 1 und die Tabelle 1 sind den Seiten 883 und 884 entnommen. Die präsentierten Ergebnisse sind Teil der zweiten Phase einer umfassenderen Studie. Die Ergebnisse von Phase I sind unter folgenden Titel veröffentlicht worden: Cole, St., / Rubin, L., / Cole, J.R., Peer Review in the National Science Foundation: Phase I of a Study. Washington, D. C.: National Academy of Sciences 1978.

23 Einen guten Überblick geben die folgenden Arbeiten: Cicchetti, D.V., The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. – In: The Behavioral and Brain Sciences. 14(1991), S. 119 – 135; Companario, J.M., Peer review for journals as it stands today, 2 Teile. – In: Science Communication. 19(1998), S. 181 – 211, S. 277 – 306; Fröhlich, G., Anonyme Kritik: Peer Review auf dem Prüfstand der Wissenschaftsforschung. – In: Eveline Pipp (Hrsg.), Drehscheibe E-Mitteuropa. Information: Produzenten, Vermittler, Nutzer. Die gemeinsame Zukunft (Biblos-Schriften Band 173). Wien: Phoibos 2002. S. 129 – 146.

- Die Uneinigkeit erstreckt sich über das gesamte Qualitätsspektrum<sup>25</sup> und umfasst mehr oder weniger alle Eigenschaften des zu beurteilenden Artikels.<sup>26</sup>
  - Sie finden oft die für die Bewertung der wissenschaftlichen Qualität entscheidenden Fehler in Manuskripten nicht.<sup>27</sup>
  - Sie veranlassen die Autoren im Revisionsverfahren zu Verbesserungen<sup>28</sup>, zwingen sie zuweilen aber auch zu „Korrekturen“, die von den Autoren
- 24 Wie häufig Gutachter bei der Bewertung von Manuskripten oder Forschungsanträgen übereinstimmen, hängt offenbar auch vom untersuchten Fach und vielleicht auch vom untersuchten Zeitraum ab. Merton und Zuckerman zitieren in ihrer Arbeit über „Institutionalized patterns of evaluation in science“ (in: Merton, R.K., *The Sociology of Science*. Chicago & London: University of Chicago Press 1974. S. 461.) empirische Untersuchungen zu drei Zeitschriften aus den genannten Gebieten, nach denen sich folgende Übereinstimmungsraten ergeben: Physik 93%, Sozialwissenschaft 73%, Biomedizin 65 – 75%. Hierbei ist zu beachten, dass eine Übereinstimmung von 50% als Erwartungswert gilt, wenn die beiden Urteile nicht korreliert sind. Andere Untersuchungen zeigen, dass bei der Messung von Übereinstimmung nach positiven und negativen Voten differenziert werden muss. In der Zeitschrift „Angewandte Chemie“ betrug z.B. die globale Übereinstimmung über die Annahme von Manuskripten (bei allerdings sehr unterschiedlicher Bewertung im Detail) 83% (253 von 304 Fällen) während im Falle einer Ablehnung nur eine Übereinstimmung von 45.4% (40 von 88 Fällen) bestand (Daniel, H.-D., *Guardians of Science*. Weinheim: VCH 1993. Table 10, S. 27; Cicchetti, D.V., *Referees, editors, and publication practices*. – In: *Science and Engineering Ethics*. 3(1997), S. 54f.) Cicchetti weist darauf hin, dass das Verhältnis im Bereich Verhaltenswissenschaft und Medizin tendenziell umgekehrt ist (70–78% Übereinstimmung bei Ablehnung vs. 45–61% bei Annahme) (Cicchetti, D.V., op. cit., 55). Insgesamt sind diese Daten schwer vergleichbar und dem Anschein nach nicht völlig konsistent. Das könnte bedeuten, dass es Faktoren gibt, die in ihnen nicht erfasst sind. Es ist bekannt, dass das Ausmaß des Konsenses innerhalb von Disziplinen keine Konstante ist, sondern je nach „paradigmatischer Phase“ starken Schwankungen unterliegt (vgl. Kuhn, T.S., *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt: Suhrkamp 1967). Man müsste also im Design jeder derartigen Studie die „Phase“ des im Untersuchungsbereich (untersuchtes Spezialgebiet, disziplinäre Abdeckung der jeweiligen Zeitschrift, etc.) gerade dominierenden „Paradigmas“ kontrollieren. Eine weitere nicht zu unterschätzende und daher unbedingt zu kontrollierende Variable ist die „Politik“ der Herausgeber von Zeitschriften oder der Manager von Vergabegremien der Drittmittelgeber. Je nachdem, ob man eine rein technisch orientierte („positivistische“) oder eine grundlagenbezogene und debattenorientierte Beurteilung wünscht, werden sich völlig unterschiedliche Grade der Gutachterübereinstimmung ergeben.
- 25 Aus einer Studie, die Companario an sogenannten „Citation Classics“ durchgeführt hat, ging hervor, dass etwa 6% dieser meistzitierten Arbeiten (nur eine von 5000 Arbeiten erhält den Status eines „Citation Classic“) nach Angaben der Autoren erst nach größeren Problemen mit den Gutachtern publiziert werden konnten. Sechs Prozent klingt beruhigend, aber man muss berücksichtigen, dass es sich dabei um naturwissenschaftliche Arbeiten handelt, bei denen die durchschnittliche Annahmequote bis zu 80% beträgt (Companario, J.M., *Consolation for the scientist: Sometimes it is hard to publish papers that are later highly-cited*. – In: *Social Studies of Science*. 23(1993), S. 342 – 362).

- her als taktische Kompromisse oder gar als „Verschlimmbesserungen“ empfunden werden.<sup>29</sup>
- Sie bevorzugen Artikel mit positiven und konventionellen gegenüber Artikeln mit kontroversen Ergebnissen (confirmation bias<sup>30</sup>) und sie
- 26 Lindsey, D., Precision in the manuscript review process: Hargens and Herting revisited. – In: *Scientometrics*. 22(1991). S. 313ff; Gottfredson, S.D., Evaluating psychological research reports. dimensions, reliability, and correlates of quality judgements. – In: *American Psychologist*. (1978), S. 920 – 934. Vermutungen, dass ablehnende Urteile mit größerer Einmütigkeit getroffen würden als zustimmende oder ambivalente, scheinen auf einen statistischen Scheineffekt zurückzugehen, der auf der Verengung der Untersuchung auf Zeitschriften mit hohen Ablehnungsquoten beruht. Kontrolliert man in den bisherigen Untersuchungen die Ablehnungsquote, dann verschwindet der Effekt zumeist. Vgl. dazu Demorest, M.E., Different rates of agreement on acceptance and rejection: A statistical artifact? – In: *The Behavioral and Brain Sciences*. 14(1991). S. 144f. Letzten Endes bleibt dies aber eine Frage, die nur empirisch und fallweise zu entscheiden ist.
- 27 Vgl. Baxt, W.G., et al., Who reviews the reviewers? – In: *Annals of Emergency Medicine*, Part 1. 32(1998)3, S. 310 – 317. Aus dem Abstract: „(...) the use of a preconceived manuscript into which purposeful errors are placed may be a viable approach to evaluate reviewer performance. Peer reviewers in this study failed to identify two thirds of the major errors in such a manuscript (...) Results: The manuscript was sent to 262 reviewers; 203 (78%) reviews were returned. One-hundred ninety-three reviewers recommended a disposition for the manuscript: 15 recommended acceptance, 117 rejection, and 67 revision. The 15 who recommended acceptance identified 17.3% (...) of the major and 11.8% (...) of the minor errors. The 117 who recommended rejection identified 39.1% (...) of the major and 25.2% (...) of the minor errors. The 67 who recommended revision identified 29.6% (...) of the major and 22.0% (...) of the minor errors... Sixty-eight percent of the reviewers did not realize that the conclusions of the work were not supported by the results.” (310)
- 28 Armstrong, J.S., Peer review for journals: Evidence on quality control, fairness, and Innovation. – In: *Science and Engineering Ethics*. 3(1997), S. 67; Krampen, G., / Montada, L., Peer reviews als Instrumente der Wissenschaftsevaluation in der Psychologie sowie der Fach- und Wissenschaftspolitik. – In: dies., *Wissenschaftsforschung in der Psychologie*. Göttingen: Hogrefe 2002, S. 57. Von 265 Befragten äußerte in der zuletzt zitierten Studie immerhin die Hälfte, dass das Gutachten zu einem eingereichten Artikel „sinnvolle und praktikable Änderungsvorschläge“ oder „bedeutsame Kritik/wichtige Anregungen“ enthielt. Etwa jeweils ein Viertel der Befragten gaben aber auch an, dass sich „verschiedene Reviews widersprachen“ und dass das Gutachten ein „fachliches Anliegen des Gutachters selbst“ wiedergab. Harschere Kritik äußerten die, die kundtaten, die „Review enthielt unsinnige, falsche Änderungsvorschläge“ (15.2%), „enthält innere Widersprüche“ (12.4%), der „Reviewer hatte wenig Kenntnisse in diesem Bereich“ (12.2%) oder er habe das „Manuskript nicht verstanden“ (9.5%) (a.a.O.).
- 29 Bradley, J.V., Pernicious publication practices. – In: *Bulletin of the Psychonomic Society*. 18(1981), S. 31 – 34. Aus dem Abstract: „In the peer review of their latest revised and published article, 76% encountered pressure to conform to the strictly subjective preferences of the reviewers, 73% encountered false criticism (and 8% made changes in the article to conform to reviewers' comments they knew to be wrong), 67% encountered inferior expertise, 60% encountered concentration upon trivia, 43% encountered treatment by referees as inferior, and 40% encountered careless reading by referees.” (S. 31)

benachteiligen ebenfalls Manuskripte, in denen eine gegenwärtig geschätzte Hypothese falsifiziert wird.<sup>31</sup>

- Sie lehnen innovative Artikel häufiger ab als konservative, obwohl in der Rhetorik „Originalität“ hochgehalten und gefordert wird.<sup>32</sup>

Eine Teilerklärung für das im letzten Ergebnis enthaltene Paradoxon könnte darin liegen, dass Originalität von den Gutachtern sehr häufig nicht als solche erkannt wird<sup>33</sup> oder primär als Ablehnung des gegenwärtigen Forschungskonsenses wahrgenommen wird. Offenbar ziehen die betreffenden Gutachter die Grenze zwischen solider „Originalität“ bzw. „Innovation“ und geistigem Abenteuerum anders als zum Beispiel Wissenschaftshistoriker oder Wissenschaftssoziologen dies im Rückblick tun. Hinterher ist man natürlich immer schlauer. Viele bahnbrechende Arbeiten fallen dem – zuweilen im Gewande der Solidität daherkommenen – Konservatismus der „Peers“ zum Opfer und können (wenn überhaupt)

- 30 Eine von Heinz Sahner unternommene Analyse von Zeitschriften ergab, dass von den sozialwissenschaftlichen Studien, die explizit der Prüfung einer Hypothese gewidmet waren, 75% in ihrer Bestätigung und nur 25% in ihrer Widerlegung resultierten. (Sahner, H., *Veröffentlichte empirische Sozialforschung: Eine Kumulation von Artefakten? Eine Analyse von Periodika*. – In: *Zeitschrift für Soziologie*. 8(1979)3, S. 267 – 278; vgl. auch Armstrong, J.S., *Peer review for journals*, op. cit., S. 71, sowie die Arbeiten von Michael Mahoney, insbesondere sein Buch: *Scientists as Subject: The psychological imperative*. Ballinger 1976.)
- 31 Armstrong, J.S., *Peer Review for Journals*, op. cit., S. 71. Beispiele findet man in: Martin, B., *Suppression Stories*. Wollongong 1997, vor allem Kap. 5: <http://www.uow.edu.au/arts/sts/bmartin/dissent/documents/>.
- 32 Vgl. Spier, R.E., *Peer review and innovation*. – In: *Science and Engineering Ethics*. 8(2002), S. 99 – 108; Armstrong, *Peer Review for Journals*, op. cit., S. 70f.; Ruderfer, M., *The fallacy of peer review – Judgment without science and a case history*. – In: *Speculations in Science and Technology*. 3(1980), S. 533 – 562. Bei den NIH (National Institutes of Health) der USA bildet „innovation (novel concepts, approaches, methods, challenge to existing paradigms“) eine der fünf Dimensionen, auf die hin Gutachter einen Antrag prüfen sollen. Dagegen fordern dies die NSF (National Science Foundation) der USA, aber auch die staatliche Forschungsförderung Großbritanniens nicht explizit. Inzwischen gibt es in den USA eine Initiative, die von den NIH abgelehnten Projektanträge im World Wide Web zu veröffentlichen, um potentielle Geldgeber aufmerksam zu machen. Das zentrale Argument des Begründers dieser Initiative, George M. Kurzon, ist, die NIH sei „a very efficient screening tool to screen out innovation“ (Peg Brickley, *Giving grant proposals a second chance*. – In: *The Scientist*, March 18, 2003; <http://www.biomedcentral.com/news/20030318/>). Als „worst offences“ des Peer Review Systems bezeichnen Rustum Roy und James Ashburn von der Pennsylvania State University „the enormous waste of scientists’ time, and the absolute, ineluctable bias against innovation“. (The perils of peer review. – In: *Nature*. 414(2001), S. 394.) Dies sind nur wenige Stimmen von vielen, die gleichlautende Kritik äußern.
- 33 Vgl. Cicchetti, D.V., *Referees, editors, and publication practices: Improving the reliability and usefulness of the peer review system*. – In: *Science and Engineering Ethics*. 3(1997), S. 51 – 62.

zunächst nur in Zeitschriften zweiter oder dritter Wahl publiziert werden.<sup>34</sup> Wie viele unveröffentlicht bleiben, wissen wir nicht, weil niemand darüber Buch führt.

Dies ist ein Punkt von strategischer Bedeutung für die Funktionsweise des Peer Review Verfahrens, auf den man bisher nicht deutlich genug geachtet hat, obwohl einige Analytiker und Wissenschaftler darauf hingewiesen haben.<sup>35</sup> Armstrong weist in seiner Arbeit aus dem Jahr 1997 darauf hin, dass in einem Klassifikationsversuch der Arbeiten zum Peer Review System, der (ebenfalls 1997) von Stamps unternommen wurde, das Thema Innovation nur an 12. Position – und hier noch unter der Überschrift „Konservatismus“ – auftaucht. Inzwischen haben Drittmittelförderer wie die amerikanischen NIH (National Institutes of Health) auf diese Vorwürfe reagiert und versuchen zumindest, die Sensibilität von Antragstellern und Gutachtern gegenüber dem Thema Innovation zu erhöhen, indem sie entsprechende Hinweise in ihre Kriterienkataloge aufnehmen.

Die bisher präsentierten Ergebnisse und Überlegungen sind um so alarmierender, als die Gutachtersysteme von Fachzeitschriften und Drittmittelgebern prinzipiell dem gleichen Muster folgen, obwohl es durchaus Unterschiede im einzelnen gibt.<sup>36</sup> Aus den Ergebnissen der empirischen Untersuchungen können wir zusammenfassend die folgenden vorläufigen Schlussfolgerungen ziehen.

Dieses System

- verfügt entweder nicht über klare Maßstäbe für wissenschaftliche Qualität oder weiß sie nicht konsistent anzuwenden,<sup>37</sup>
- belohnt Konformität oder taktische Allianzen mit bestimmten theoretischen

34 Armstrong, J.S., *Peer Review for Journals*, op. cit., S. 71.

35 Günter Blobel: „If you can predict what you're going to do for five years, its probably going to be bad.“ (Goodman, B., *Observers fear funding practices may spell the Death of innovative grant proposals*, June 1995 <http://www.the-scientist.library>.) Erwin Chargaff bemerkt: „The so-called advance of science rests, in most cases, on two kinds of observation: predictable and unpredictable. The major part is of the first kind, predictable; it grows out of the accumulated body of accepted knowledge, and these observations can very well be made by teams or at least by several people in collaboration. The much rarer kind, the unpredictable observations, are the only ones deserving the name of discovery, and they are always due to a single person.[...] The trend is all toward the creation of very large scientific conglomerates in which, under the leadership of men with managerial qualifications, the predictable will be discovered in ton lots. [...] The frightening waste of resources will become evident to anybody who considers how little of value the orgy of goal directedness has actually produced. One could, in fact, argue that our scheme of research support has much more harmed than helped the scientific growth of the individual.“ (Chargaff, E., *In praise of smallness*. – In: *Perspectives in Biology and Medicine*. 23 (1980), S. 37)

36 Die Vermutung liegt nahe – dies sei nebenbei bemerkt –, dass die auf der Basis eines solchen Systems eingeworbenen Drittmittel auch kein zuverlässiges Maß der Forschungsleistung einer Universität sein können.

schen und methodischen Ausrichtungen, wobei diese Konformität nicht total sein darf, sondern individuelle Nuancierungen aufweisen sollte,<sup>38</sup>

- bevorzugt bestimmte Paradigmen, Themen und Argumentationsfiguren,<sup>39</sup>
- prämiiert die soziale Macht und den Bekanntheitsgrad von Antragstellern und Institutionen,<sup>40</sup>

- 37 Aus der unten diskutierten Studie von Gottfredson scheint sich zu ergeben, dass das Hauptproblem in der praktischen Anwendung eines im großen und ganzen nicht besonders strittigen Kriterienkatalogs besteht. (Vgl. auch Neidhardt, E., Selbststeuerung in der Forschungsförderung. Das Gutachterwesen der DFG. Opladen: Westdeutscher Verlag 1988. Tab. A12, S. 148.) Historisch gesehen sind aber auch viele Kriterien variabel.
- 38 Die Hauptfunktion dieser Selbsteinordnung besteht darin, die „Anschlussfähigkeit“ der eigenen Forschungen zu demonstrieren und die Gutachter davon zu überzeugen, dass der Autor oder Antragsteller kein intellektueller Hasardeur ist, sondern sich als Teil des Gemeinschaftsunternehmens „Wissenschaft“ sieht. Soziologisch gesehen, handelt es sich um einen Akt der sozialen Kontrolle im Wissenschaftssystem, der in Gestalt einer die Folgen der Normverletzung antizipierenden Selbstkontrolle wirksam wird.
- 39 Ein Vergleich der bei der *Zeitschrift für Soziologie* zwischen 1972 und 1980 eingereichten Artikel mit den von dieser Zeitschrift publizierten Artikeln zeigt, dass die Publikationschance sehr stark von der theoretischen Orientierung des Artikels abhing. Die durchschnittliche „Diffusionschance“ von Artikeln aus dem Umkreis des Funktionalismus war beispielsweise elfmal höher als die von Artikeln aus dem Umkreis der Kritischen Theorie und fünfmal höher als die von Artikeln aus dem Bereich des Symbolischen Interaktionalismus. (vgl. Sahner, H., Zur Selektivität von Herausgebern: Eine Input-output-Analyse der „Zeitschrift für Soziologie“. – In: *Zeitschrift für Soziologie*. 11(1982)1, S. 88, Tabelle 4.) Sahner vermutet, dass Zeitschriften mit Begutachtungssystem undurchlässiger für Innovationen sind als solche, die den Herausgebern die Entscheidung über Manuskripte überlassen. Insgesamt benachteiligt dieses Publikationssystem vor allem solche theoretische Ausrichtungen, die nur mangelhaft durch Herausgeber- oder Gutachterpräferenzen repräsentiert sind.
- 40 Vordergründig dagegen steht die Studie von Abrams (Anmerkung 1), die am Entscheidungsverfahren der amerikanischen National Science Foundation gerade kritisiert, dass die bisherige Erfolgs- oder Mißerfolgsgeschichte der Antragsteller im Beurteilungsverfahren keine Rolle spielt. Die Argumentation von Abrams wird gestützt durch eine Stellungnahme von Rosalyn Yalow (Nobelpreis für Medizin) und durch die „Ortega-Hypothese“ der Soziologen Jonathan R. Cole und Stephen Cole (*Social Stratification in Science*. Chicago & London: University of Chicago Press 1973, Kap. 8). Allerdings legen Abrams, Yalow und Cole & Cole einen etwas anderen Elitenbegriff zugrunde als den in unserer Aufzählung benutzten. Ein Widerspruch zwischen den Ansichten dieser Autoren und unserer obigen Behauptung ergibt sich erst, wenn man davon ausgeht, dass die von ihnen gemeinte *creative Leistungselite* mit der von uns angesprochenen *sozialen Prestigeelite* identisch ist. Historisch betrachtet, ist das selten der Fall, wenngleich es zumeist eine Schnittmenge gibt. Der Unterschied zwischen beiden Gruppen besteht darin, dass die Reputation der *sozialen Prestigeelite* auf Leistungen beruht, die *in der Vergangenheit* liegen, während die Reputation der  *kreativen Leistungselite* eine schwankende Größe darstellt, die auf Erwartungen beruht, über deren Zuverlässigkeit erst die *Zukunft* entscheiden kann.

- scheut das Risiko und benachteiligt die innovativen, explorativen, disziplin-übergreifenden, spekulativen Projekte gegenüber jenen, die aufgrund ihrer Einordnungsfähigkeit in einen klaren methodischen und theoretischen Rahmen zwar nur kleine, dafür aber sichere wissenschaftliche Erträge bringen werden.

Einer der Erforscher des Peer Review Systems, J. Scott Armstrong, der zugleich Mitherausgeber des *Journal of Forecasting* ist, hat auf der Grundlage seiner Erfahrungen und Erkenntnisse eine ironisierende „Autorenformel“ vorgestellt, mit der potentielle Autoren ihre Chancen verbessern und die Annahme ihrer Manuskripte deutlich beschleunigen könnten. Sie lautet: „Authors should: (1) *not* pick an important problem, (2) *not* challenge existing beliefs, (3) *not* obtain surprising results, (4) *not* use simple methods, (5) *not* provide full disclosure, and (6) *not* write clearly.“<sup>41</sup> Was ansonsten wie eine Persiflage wirken könnte, gewinnt vor dem Hintergrund der zitierten empirischen Ergebnisse und unserer Schlussfolgerungen aus ihnen geradezu den Status eines wissenschaftlichen Überlebensrezepts.

Der Befund ist relativ klar. Vielleicht war nichts anderes zu erwarten. Das Peer Review System beruht auf den Wertungen von Mitgliedern der wissenschaftlichen Gemeinschaft, also des – wie man sagen könnte – „Otto-Normalkonsumenten und -produzenten“ von Wissenschaft. Das System beurteilt sich selbst. Dabei kann man aus Gründen der Praktikabilität für den Einzelfall keinen repräsentativen Querschnitt, sondern in der Regel nur zwei oder drei Einzelstimmen heranziehen. Dass das Ergebnis angesichts der Vielstimmigkeit des Wissenschaftsbetriebs nur eine selektive Kakophonie sein kann, erscheint vielleicht nicht verwunderlich. Selbst wenn es möglich wäre, über jedes Manuskript eine repräsentative Auswahl aus der Gemeinschaft der Wissenschaftler abstimmen zu lassen, würde sich die Güte des Verfahrens nicht unbedingt sprunghaft verbessern.

## 2. *Sichert die Verbesserung der Reliabilität des Peer Review Verfahrens seine Validität?*

Domenic Cicchetti hat darauf hingewiesen, dass eine Verbesserung der Objektivität und der Reliabilität des Peer Review Verfahrens nicht unbedingt auch seine Validität erhöhen würde. „Two independent reviewers and the editor may all agree that a manuscript is not worth publishing. Yet, further developments in the

41 Armstrong, J.S., Barriers to scientific contributions: The author's formular. – In: *The Behavioral and Brain Sciences*. 5(1982), S. 197. Auch dem letzten Punkt liegt ein Versuch zugrunde. Von zwei inhaltsgleichen Manuskripten wurde das als besser bewertet, das seinen Inhalt auf sprachlich kompliziertere Weise darstellt.

field may provide scientific evidence that indicates that their decisions were incorrect.”<sup>42</sup> Cicchetti zitiert Beispiele hierfür. Als Grund dafür, warum ein heute erzielter Konsens kein zuverlässiger Indikator für die Richtigkeit der getroffenen Entscheidung ist, nennt Ronald N. Kostoff den „Pied Piper Effect“. Dieser nach dem „Rattenfänger von Hameln“ benannte Effekt war von Kostoff ursprünglich für die Interpretation von Zitationen von Zeitschriftenaufsätzen definiert worden, ist jedoch, wie der Autor bemerkt, „applicable to any conclusion resulting from any type of peer review as well: journal, proposal, program.“<sup>43</sup> Aus diesem Grund ist es auch keineswegs ein Anlass zur Genugtuung, wenn einige Studien zu anderen Disziplinen als der Psychologie zu etwas besseren Resultaten kommen als Peters und Ceci. Eine nähere Betrachtung zeigt, dass diese Ergebnisse oft nicht auf besseren Zahlen, sondern auf ihrer optimistischeren Interpretation beruhen.

Unter diesen Studien befindet sich eine Monographie von H.-D. Daniel, die 1993 unter dem (ernstgemeinten) Titel „Guardians of Science“<sup>44</sup> erschienen ist. Daniel hat die Entscheidungen der Zeitschrift *Angewandte Chemie* über 449 Publikationsangebote des Jahres für die Kategorie der „Communications“ untersucht.

42 Cicchetti, D.V., Referees, editors, and publication practices, op. cit., S. 58.

43 Kostoff, R.N., Research program peer review: Principles, practices, protocols. Arlington: Office of Naval Research <http://www.dtic.mil/dtic/kostoff/Peerweb>. Das Argument von Kostoff lautet: „Assume there is a present-day mainstream approach in a specific field of research; for example, the chemical/ radiation/ surgical approach to treating cancer [...]. Assume the following hypothetical scenario: there exist alternative approaches to treatment not supported by the mainstream community; in fifty years a cure for cancer is discovered; the curative approach has nothing to do with today's mainstream research, but is perhaps a downstream derivative of today's alternative methods; it turns out that today's mainstream approach sanctioned by the mainstream medical community was completely orthogonal or even antithetical to the curative approach. Then what meaning can be ascribed to research papers in cancer today which are highly cited for supposedly positive reasons? In this case, a paper's high citations are a measure of the extent to which the paper's author has persuaded the research community that the research direction contained in his paper is the correct one, and not a measure of the intrinsic correctness of the research direction. It is analogous to firing a missile accurately at the wrong target. In fact, the high citations may reflect the deliberate desire of a closed research community (the author and the citers) to persuade a larger community (which could include politicians and other resource allocators) that the research direction is the correct one. This is the 'Pied Piper' effect. The large number of citations in the above hypothetical medical example becomes a measure of the extent of the problem, the extent of the diversion from the correct path, not the extent of progress toward the solution. The 'Pied Piper' effect is a key reason why, especially in the case of revolutionary research, citations and other quantitative measures must be part of and subordinate to a broadly constituted peer review in any credible evaluation and assessment of research impact and quality” (a.a.O.). Das Argument lässt sich zwanglos auf den Entscheidungsprozess in jeglicher wissenschaftlicher Begutachtung anwenden.

44 Vgl. Daniel, H.-D., Guardians of Science. Fairness and Reliability in Peer Review, a.a.O.

Seine Daten stimmen – was die reinen Zahlen angeht – im wesentlichen mit den Ergebnissen anderer Untersuchungen überein. Was die Reliabilität des Prozesses, also die Übereinstimmung zwischen verschiedenen Gutachtern, angeht, fielen die Kappa- und Intraklassen-Koeffizienten bei den einzelnen Bewertungsdimensionen in den Bereich zwischen 0.12 und 0.23 (vgl. Tabelle 3). „From a statistical standpoint, the observed extent of referee agreement must be regarded as rather unsatisfying.“<sup>45</sup>

Tabelle 3: *Reliabilitätsmaße der Begutachtungen für die Zeitschrift „Angewandte Chemie“*

Fragen	Übereinstimmende Gutachterpaare	Tatsächliche Übereinstimmung	Erwartungswert für zufällige Übereinstimmung	Cohen's Kappa-Koeffizient
Are the contents of the manuscript of wide and general interest? (Yes/No)	204	0,65	0,54	0,23
Are the contents of the manuscript of extraordinary but special interest? (Yes/No)	107	0,64	0,58	0,12
Do the data obtained by experiment or calculation verify the hypotheses and conclusions? (Yes/No)	296	0,82	0,78	0,17
Is the length of the manuscript appropriate to its contents? (Yes/No)	309	0,67	0,62	0,13
The form of the manuscript (text, figures, tables, nomenclature etc.) is beyond reproach? (Yes/No)	297	0,65	0,60	0,12
Do you recommend acceptance of the Communication? (Yes/No) – alle Kategorien	392	0,38	0,29	0,14 <sup>a b</sup>

a. Cohen's Kappa-Koeffizient = 0,20

b. ANOVA Intraklassen-Korrelationskoeffizient = 0,25

Daniel glaubt jedoch, dass diese Zahlen das wahre Ausmaß der Übereinstimmung nicht korrekt wiedergeben. „The true level of existing agreement is systematically underestimated.“<sup>46</sup> Zum einen würde es sich oft nur um kleinere Diskrepanzen handeln, die die Frage, ob eine Arbeit grundsätzlich publikationswürdig sei, nicht tangieren, zum anderen gingen Bewertungsunterschiede oft nicht auf reale Meinungsunterschiede, sondern auch auf „dislocational compo-

45 Daniel, H.-D., Guardians of Science. op. cit., 71; folgende Tabelle (Table 3) auf S. 24.

46 Daniel, H.-D., Guardians of Science. op. cit., S. 72.

Tabelle 4: *Neu berechnete Konsensmaße abgelehnter und akzeptierter „communications“ bei zwei Gutachtern (in Prozent)*

	Degree of consensus <sup>a</sup>			
	++	+-	+-	--
Accepted communications (N = 286)	37	43	14	7
Rejected communications (N = 103)	43	25	25	7
All communications (N = 392)	38	38	17	7

- a. ++ : Both referees offered identical recommendations  
 +- : Referees recommendations differed by one category  
 +--: Referees recommendations differed by two categories  
 -- : Referees disagreed completely

nents“ und „differences in the frames of reference“<sup>47</sup> zurück. Wenn man diese Faktoren berücksichtigt und alle Differenzen ignoriert, die nur eine Kategorie auseinander liegen, dann käme man auf eine weit höhere, im Bereich von 0.67 liegende Gutachterübereinstimmung (vgl. Tabelle 4).

Über die Bedeutung der von Daniel als unwesentlich herausgerechneten Differenzen kann man natürlich streiten.<sup>48</sup> Man kann mit guten Gründen argumentieren, dass eine Verschiebung des Bezugsrahmens eine wichtige Perspektivenänderung des Faches anzeigt, die für die Bewertung einer großen Anzahl von Arbeiten von prinzipieller Bedeutung ist. So gesehen, wäre eine perspektivische Differenz zwischen Gutachtern nicht herauszurechnen, sondern besonders zu betonen, da sie von höherer Wertigkeit als eine Differenz im Detail ist. Man kann auch darüber streiten, ob die ursprünglichen Zahlen das wahre Ausmaß der Diskrepanz in Wirklichkeit nicht vergrößern, sondern verkleinern. Die Gutachter können sich nämlich aus ganz unterschiedlichen Gründen auf ein bestimmtes Urteil festlegen. Wenn zwei Gutachter auf „Akzeptieren nach größeren Veränderungen“ plädieren, dann ist damit nicht gesagt, dass sie identische Änderungswünsche haben. Auch kompromisslose Ablehnungen oder sofortige Zusagen können auf un-

47 Daniel, H.-D., *Guardians of Science*. op. cit.

48 Ein von Daniel in seiner Auswirkung nicht untersuchter Faktor ist die Verteilung der Gutachter auf die verschiedenen Eingaben. Aus seiner Tabelle 5 (Daniel, H.-D., *Guardians of Science*. op. cit., 17) geht hervor, dass die zehn Gutachter, die am häufigsten befragt wurden, etwa ebenso viele Gutachten verfassten wie die 152 Gutachter, die nur einmal um eine Stellungnahme gebeten wurden. Es handelt sich dabei um eine typische Lotka-Verteilung, die die Frage provoziert, wieviel Prozent an Übereinstimmung oder Diskrepanz im Sample durch eine eventuelle und faktisch vermutlich kaum vermeidbare Häufung von Gutachter-Paaren erzeugt wird.

terschiedlichen Gründen beruhen, die sich bei näherer Analyse vielleicht sogar widersprechen. Obwohl Daniels Vorhaben, die Zahlen auf ihre Bedeutung zu hinterfragen, vollkommen legitim ist, braucht man differenziertere Erhebungen und neue theoretische Überlegungen, bevor man abschätzen kann, ob eine tiefere Interpretation der Daten ein höheres Maß an Übereinstimmung oder an Diskrepanz ergeben wird.

Die Studie von Daniel ist aber noch aus einem anderen Grund bemerkenswert. Der Autor versucht herauszufinden, ob dem vermuteten Maß an Reliabilität der Gutachterentscheidung auch ein entsprechendes Maß an Validität entspricht. Als Maß der Validität nimmt Daniel die Zahl der Zitationen, die ein Artikel fünf Jahre nach Publikation erzielt hat.<sup>49</sup> Daniel gelingt es nachzuweisen, dass die von der Zeitschrift *Angewandte Chemie* akzeptierten 323 Artikel in den folgenden fünf Jahren im Schnitt 11.5 Zitationen erhielten, während die von *Angewandte Chemie* abgelehnten und dann in anderen Zeitschriften publizierten 88 Artikel im Schnitt nur 6 Zitationen erhielten. Dies interpretiert Daniel als Indiz dafür, dass die Entscheidungen der Zeitschrift *Angewandte Chemie* ein beträchtliches Maß an Validität aufwiesen.

Diese Schlussfolgerung bedarf der Überprüfung. Zunächst setzt sie voraus, dass Zitationen ein gültiges Maß für die Validität der von Gutachtern und Herausgebern abgegebenen Bewertungen sind. Ganz abgesehen von der Frage, ob Zitationen als Maß für wissenschaftliche Qualität oder Leistung interpretiert werden dürfen,<sup>50</sup> kann die Unabhängigkeit der beiden Beurteilungsmaßstäbe angezweifelt werden. Anders gesagt, man kann vermuten, dass den kurz nach Veröffentlichung einer Publikation erzielten Zitationen ähnliche Maßstäbe zu Grunde liegen, wie sie auch von Gutachtern und Herausgebern verwendet wurden.<sup>51</sup> Erst in längerfristiger Betrachtung werden sich die beiden Messlatten so weit entkoppelt haben, dass keine Autokorrelationen mehr zu erwarten sind.

Das Hauptargument gegen Daniels Anscheinsbeweis ist aber ein anderes. Die angeführten Zahlen sind statistische Durchschnitte. Betrachtet man die Vertei-

49 Genauer gesagt, müsste man einschränkend auf die Zahl der Zitationen in den Datenbanken des ISI hinweisen. Es ist bekannt, dass diese Zahl nicht die wahren Zitationen wiedergibt, sondern sie – je nach Publikationssprache, Publikationsland oder Disziplin – mehr oder weniger stark verzerrt, von nachgewiesenen Kodierungsfehlern ganz abgesehen.

50 Dem widersprach bereits der Erfinder des Citation Index, Eugene Garfield. „People talk about citation counts being a measure of the 'importance' or 'impact' of scientific work, but those who are knowledgeable about the subject use these words in a very pragmatic sense; what they really are talking about is utility. A highly cited work is one that has been found to be useful by a relatively large number of people, or in a relatively large number of experiments.” (Garfield, E., *Is citation analysis a legitimate evaluation tool.* – In: *Scientometrics*. 1(1979), S. 363).

51 Meines Wissens hat Daniel nicht nachgeprüft, von wem die Zitationen jeweils stammen.

lung im einzelnen, dann wird ersichtlich, dass sie einen großen Überlappungsbe-  
reich haben. Anders gesagt, viele der von der Zeitschrift *Angewandte Chemie*  
abgelehnten Publikationen haben höhere Zitationsziffern erreicht als viele der an-  
genommenen. Von den angenommenen Artikeln haben ca. 30% weniger als 7  
Zitationen erhalten, während von den abgelehnten Artikeln fast 40% 7 und  
mehr Zitationen erzielten.<sup>52</sup>

Der partielle Konsens spricht für ein entsprechendes Maß an Objektivität und  
Reliabilität des Verfahrens, sagt aber nichts über seine Validität aus. Validität ist  
die am schwierigsten zu sichernde Qualität von Untersuchungen oder Verfahren.  
Wie kann man sie messen? Eine eher resignative Lösung bestünde darin, der Ge-  
schichte die Entscheidung zu überlassen. Wir wissen heute, dass die Kritiker un-  
ter den Peers von Kopernikus, Kepler, Galilei, Newton, Mayer, Darwin, Mendel,  
Einstein, Heisenberg usw., nicht immer im Detail, aber im wesentlichen Unrecht  
hatten.<sup>53</sup> Aber dies sind Figuren der Vergangenheit. Gibt es eine Möglichkeit,  
dieses Wissen in einer methodisch operationalisierbaren (also intersubjektiv nach-  
vollziehbaren) Weise schon früher zu erlangen, also die Heisenbergs, Mendels und  
Einsteins in Gestalt ihrer geistigen Produkte so zeitig zu erkennen, dass man ihre  
Arbeit optimal fördern kann?

Eine der wenigen Arbeiten, die für die Beantwortung dieser Frage relevant er-  
scheinen, ist die groß angelegte Studie von Stephen Gottfredson. In dieser im  
Oktober 1978 im *American Psychologist* (S. 920ff) erschienenen Studie legte der  
Autor Daten vor, die die Urteile von Experten über zehn Jahre zuvor (1968) er-  
schienene psychologische Artikel mit den in den acht Jahren nach Publikation er-  
haltenen Zitationen in Beziehung setzen. Gottfredson hat über eine Befragung  
der Herausgeber und Redakteure von neun amerikanischen psychologischen  
Fachzeitschriften faktorenanalytisch Cluster von Merkmalen isoliert, die bei der  
Entscheidung über die Publikationswürdigkeit (Qualität und „impact“) von Ma-

52 Vgl. Daniel, H.-D., *Guardians of Science*. op. cit., Figure 7, S. 53. Dies muss wiederum nicht  
bedeuten, dass diese 40% zurückgewiesenen Artikel in Wahrheit *besser* sind als die angenom-  
menen 30%. Es kann auch heißen, dass die Zielgruppen für die abgelehnten Artikel zum Teil eher  
anderen Zeitschriften als der Zeitschrift „*Angewandte Chemie*“ zuzuordnen waren. Wenn diese  
Zeitschriften ein spezialisierteres Publikum bedienen, dann wird sich dies *ceteris paribus* in  
einem niedrigeren impact-Faktor niederschlagen. Der von Daniel beabsichtigte Vergleich wird  
dadurch unsinnig. Ein korrekter Vergleich wäre nur dann möglich, wenn die abgelehnten Arti-  
kel ebenfalls in der *Angewandte Chemie* erschienen wären. Vermutlich wird sich kein Herausge-  
ber auf dieses Experiment einlassen.

53 Dieses Wissen kann allerdings nicht die Sicherheit logischer Ableitungen beanspruchen. Prinzi-  
piell wäre es möglich, dass die zukünftige Entwicklung der Wissenschaften wichtige Aspekte  
dieser Urteile wieder in Frage stellen kann. Grundlegende Änderungen der Sichtweise sind  
ebenso wenig vorhersehbar wie ihre Folgen.

nuskripten für relevant gehalten werden. Experten wurden darauf hin (auf Vorschlag der Autoren der ausgewählten Artikel) befragt, wie sie die betreffenden Artikel vor dem Hintergrund der erhobenen Merkmale einschätzen würden. Diese Einschätzungen wurden anschließend mit Zitationsmassen korreliert.

Ein wichtiges Ergebnis der Untersuchung war, dass zwischen Psychologen, die für die neun beteiligten Fachzeitschriften begutachteten, eine „bemerkenswerte“ Übereinstimmung über die Merkmale bestand, die die eingereichten Manuskripte aufweisen oder nicht aufweisen sollten. Dies wird von Gottfredson dahingehend interpretiert, dass das Fach über gemeinsame Wertmaßstäbe verfügt. In ihrer Anwendung auf die erneut zu bewertenden Publikationen wiesen die erhaltenen Skalen – von den beiden letzten (s.u.) abgesehen – eine gute Intra-Gutachter-Übereinstimmung (0.89 bei Kombination der Skalen 1–7) auf, was man als Selbstkonsistenz interpretieren kann. Wesentlich schlechter fiel die Inter-Gutachter-Übereinstimmung (Reliabilität) aus. Sie lag – wiederum ohne die beiden letzten Skalen – bei 0.49, was bedeutet, dass 24% der bei der Beurteilung der Artikel zu erklärenden Varianz auf den wertbasierten Konsens der Gutachter zurückgeht. Dass diese Zahl etwas höher liegt als bei anderen Studien zur Reliabilität des Peer Review Systems, könnte durch zwei besondere Bedingungen der Studie erklärbar sein:

Abweichend von der Praxis vieler Herausgeber, kontroverse Gutachten einzuholen, hat sich Gottfredson nach den Empfehlungen der Autoren der zu begutachtenden Artikel gerichtet. Diese sollten Kollegen benennen, von denen sie annehmen, dass sie in der Lage wären, ihre Arbeit zu bewerten. Dass bei diesem Verfahren nicht die schärfsten Kritiker zum Zuge kommen, erscheint plausibel.

Die Bewertungen erfolgten zehn Jahre nach Veröffentlichung, also zu einem Zeitpunkt, zu dem viele strittige Fragen geklärt, Kontroversen beigelegt und der Wert der früheren Arbeiten vermutlich besser eingeschätzt werden konnte.

Überraschend fielen jedoch die daraufhin berechneten Korrelationen der neuerlichen Begutachtungen mit den inzwischen erhaltenen Zitationen eher schwach aus. Dies erlaubt den Schluss, dass die Expertenurteile nur ein schlechter Prädiktor für die erhaltenen Zitationszahlen sind, oder umgekehrt, dass Zitationsmasse kein angemessener Prädiktor für wahrgenommene Qualität und wahrgenommenen „impact“ sind. Die Tabelle 5 zeigt die Ergebnisse.

Diese Ergebnisse entsprechen den durch die bisherigen Überlegungen und Daten geweckten Erwartungen.<sup>54</sup> Die Korrelationen sind zwar besser als in ver-

54 Die Faktoren der Skala sind nur zum Teil selbsterklärend. Da die genauere Beschreibung ihrer einzelnen Komponenten zu umfangreich ist, um sie hier abzudrucken, sei der Leser auf die Originalquelle verwiesen: Gottfredson, S.D., Evaluating psychological research reports. Dimensions, reliability, and correlates of quality judgements. – In: *American Psychologist*. (1978), S. 920 – 934. Die Tabelle befindet sich auf S. 931.

Tabelle 5: *Korrelation zwischen Expertenurteilen und Zitationsmaßen*

Experts' judgments	Citation measure		
	Total citations	Total citations by others	Total review citations
Quality scale (382)	+.24	+.22	+.11
Impact scale (380)	+.37	+.36	+.16
Scale 1 – Don't's (331)	-.05	-.04	+0.03
Scale 2 – Substantive do's (335)	+0.23	+0.22	+0.15
Scale 3 – Stylistic/compositional do's (335)	+0.08	+0.07	+0.06
Scale 4 – Originality/heurism (340)	+0.15	+0.13	+0.07
Scale 5 – Trivia (339)	-.18	-.16	-.07
Scale 6 – Where do we go? (321)	+0.17	+0.17	+0.06
Scale 7 – Data grinders (339)	-.10	-.09	-.01
Scale 8 – Ho-hum research (320)	-.12	-.11	-.06
Scale 9 – Magnitude of problem/interest (319)	-.05	-.08	-.13
Scale 1 – 9 combined (340)	+0.19	+0.18	+0.10

gleichbaren Studien, bleiben aber bescheiden, die Experten können die von den Arbeiten erreichten Zitationsziffern auch im Nachhinein nicht korrekt schätzen.

Auch der Bezug auf spezielle Qualitätsmerkmale verbessert die Situation nicht. Eine Differenzierung des Samples in hochzitierte (auf oder oberhalb des Medians) und niedrigzitierte Arbeiten (unterhalb des Medians) hatte das Teilergebnis zur Folge, dass die Korrelation zwischen Expertenurteil und dem Zitationsmaß „Total Citations“ bei den niedrigzitierten Arbeiten verschwindet, während sich die Korrelation bei den hochzitierten Arbeiten auf .33 („Artikelqualität“) und .39 („Artikel-impact“) erhöht.<sup>55</sup> Dies bedeutet, dass die Gesamtkorrelation zwischen Expertenurteil und Zitationsmaßen ausschließlich auf die Korrelation innerhalb des Segments der hochzitierten Arbeiten zurückzuführen ist. Anders ausgedrückt: die Experten wissen nicht, welche der von ihnen insgesamt als hochwertig eingeschätzten Arbeiten stark zitiert, also von der wissenschaftlichen Gemeinschaft als aktuell nützlich eingeschätzt wurden, aber unter den vielzitierten („nützlichen“) Arbeiten können sie in begrenzten Umfang summarische Qualitäts- und Impactstufen identifizieren. Allerdings wissen sie wiederum nicht, mit welchen spezifischen Eigenschaften einer Arbeit diese Differenzierung zu begründen ist.

55 Das heißt, dass im ersten Fall 11% und im zweiten Fall 15% der Varianz der Zitationsziffern durch die Expertenurteile erklärt werden.

### 3. *Peer Review und innovative Forschung*

Es ist klar, dass die durch Expertenurteil oder per Zitationszählung gemessene „Nützlichkeit“ einer Arbeit zeitgebunden ist und vor dem Urteil der Wissenschaftsgeschichte oft nicht bestehen kann. Sie ist insbesondere kein Maß für die Bedeutung einer Arbeit für die Geschichte einer Disziplin, wie sie von Wissenschaftshistorikern im Rückblick und mit dem Wissen des später Geborenen geschrieben wird.

Man kann das auch etwas anders ausdrücken: Über Wahrheit – oder besser: über die Urteile, die die Wissenschaft der Zukunft fällen wird – können die Experten heute nicht per Konsensbeschluss entscheiden. Auch die wissenschaftliche Gemeinschaft kann dieses Urteil nicht in Form aktueller summarischer Ziffern der Zitation vorwegnehmen. Die Zukunft ist offen. Viele Phänomene sind unentdeckt, viele Störfaktoren und ihre Interaktionen unerforscht, viele Gesetzmäßigkeiten unerkannt. Der aktuelle wissenschaftliche Konsens ist nichts weiter als der gerade kursierende Irrtum. Niemand weiß, aus welcher Ecke der entscheidende Anstoß für den nächsten wissenschaftlichen Durchbruch kommen wird und welche Kombination von Informationen dafür entscheidend sein wird. „In fundamental science“ – so meint Erwin Chargaff nach einem langen Forscherleben – „the unpredictable happens when it is least expected. But for it to happen, there must exist the possibility of a very large number of unforeseeable free associations, of entirely unplanned collisions. The freer, the less trammled the scientist, the greater the chance of new principles being found.“<sup>56</sup>

Den kreativen Forscher, der auf Unterstützung seiner Projekte durch Drittmittel angewiesen ist, bringt das allerdings in ein schwer auflösbares Dilemma. Je innovativer und origineller seine Projektideen oder Artikel sind, desto schwerer sind die Erfolgsaussichten für jeden potentiellen Gutachter einzuschätzen. Einzelne werden ihn mögen, andere werden vehement dagegen sein. Dahinter stehen jeweils Hoffnungen oder Befürchtungen, aber niemand wird sein Urteil mit objektivierbaren Gründen untermauern können. In der Regel bedeutet das Ablehnung, denn Zeitschriften wie Drittmittelgeber scheuen das Risiko, wenn auch aus verschiedenen Gründen. Zeitschriften sind um ihre Reputation besorgt, Drittmittelgeber sind rechenschaftspflichtig und stehen heute mehr denn je unter Legitimationsdruck. Das bedeutet, dass sie im Zweifel trotz des kleineren zu erwartenden Ertrages auf sichere Projekte setzen und den möglichen Ertrag riskanter Vorhaben eher als Spekulationsgewinne einschätzen, auf den zu bauen unverantwortlich und unseriös wäre.

56 Chargaff, E., In praise of small science. – In: *Perspective in Biology of Medicine*. 23(1980), S. 53.

Dieser Zusammenhang verweist auf eine Grundschwäche der Projektforschung: Begutachtete Projektforschung kann nie „Forschung ins Blaue hinein“ sein, auf der Basis des Prinzips Hoffnung, geleitet von vagen Ideen, kühnen Spekulationen, Analogien und Metaphern. Sie braucht ein klares Ziel, einen praktikablen Plan, bewährte Methoden, einen festen Zeitrahmen und genaue Arbeitsgrundlagen, die die „vermuteten Ergebnisse“ nach Möglichkeit bereits hypothetisch vorwegnehmen.

Der Molekulargenetiker und Nobelpreisträger Joshua Lederberg hat den Widersinn dieses Verfahrens und die Kehrseite dieser Art von Forschung wie folgt beschrieben. „The implication that an investigator should ‘know what he is doing’ before being worthy of a grant flies in the face of the actual history of the most creative discovery. How would a project proposal to NSF (National Science Foundation – K. F.) have fared that looked to explore the high-temperature superconductivity of ceramics? And I will aver in retrospect about my own career since 1946 that none of my own most consequential discoveries had been telegraphed in project proposals beforehand. About the most important matters, we are always too ignorant in advance to spell out the discoveries we might make.”<sup>57</sup> Wenn die durch „Peers“ begutachtete Projektforschung zur dominierenden Form der Forschungsförderung wird, dann sind nach den gegenwärtigen Richtlinien explorative Untersuchungen ohne klare Zielsetzung aber mit der Chance fundamentaler Neuerungen nur noch insoweit möglich, als es den Projektnehmern gelingt, diese unter dem Deckmantel anderer, konventionellerer Forschungen zu verstecken – eine Form von Etikettenschwindel, die funktional für den Fortschritt der Wissenschaft werden könnte. Erfahrene Projektnehmer stellen unterdessen nur noch Anträge für Projekte, die bereits abgeschlossen sind und deren Ergebnisse sie kennen. Die Gelder verwenden sie für ein neues Projekt, von dem sie zwar vermuten, dass es vielversprechend ist, dessen wahre Bedeutung und dessen Ergebnisse sie allerdings nicht kennen können. Sie wissen, dass sie ein hohes Risiko eingehen würden, wenn sie ihre wahren Ziele in Form eines Projektantrags bekanntgeben würden.<sup>58</sup>

Man kann versuchen, eine „Skala der Originalität“ von Innovationen aufzustellen und diese mit dem Ablehnungsrisiko entsprechender Projektanträge oder Manuskripteinreichungen zu korrelieren. Forschungsarbeiten können unterschiedliche Ziele verfolgen und unterschiedliche Ansprüche erheben. Sie können zum Beispiel

- ein gängiges methodisches und begriffliches Instrumentarium zur Lösung von offenen Problemen in anerkannten Anwendungsbereichen anwenden,

57 Lederberg, J., Does scientific progress come from projects or people? – In: Garfield, E., Creativity, Delayed Recognition, and Other Essays (Essays of an Informations Scientist Vol. 12). Philadelphia: ISI Press 1991. S. 340.

- Standardverfahren auf neue Gegenstandsbereiche übertragen,
- zeigen, wie man eine Idee, einen Begriff, eine Theorie operationalisieren oder eine theoretische Größe messen könnte,
- Implikationen neuer Beobachtungen und Experimente für aktuelle Hypothesen und Theorien herausarbeiten,

58 Diese Strategie hat noch einen anderen Grund. Erfahrene Forscher wissen, dass Ideen gestohlen werden oder zumindest auf verschlungenen Wegen zu potentiellen Konkurrenten diffundieren können. Der bekannte Molekulargenetiker Bentley Glass hat dies wie folgt ausgedrückt. „What has been said about referees applies with even greater force to the scientists who sit on panels that judge the merit of research proposals made to government agencies or to foundations. The amount of confidential information directly applicable to a man's own line of work acquired in this way in the course of several years staggers the imagination. The most conscientious man in the world cannot forget all this, although he too easily forgets when and where a particular idea came to him. This information consists not only of reports or what has been done in the recent past but of what is still unpublished. It includes also the plans and protocols of work still to be performed, the truly germinal ideas that may occupy a scientist for years to come. After serving for some years on such panels I have reached the conclusion that this form of exposure is most unwise. One simply cannot any longer distinguish between what one properly knows, on the basis of published scientific information, and what one has gleaned from privileged documents. The end of this road is self-deception on the one hand, or conscious deception on the other, since in time scientists who must make research proposals learn that it is better not to reveal what they really intend to do, or to set down in plain language their choicest formulations of experimental planning, but instead to write up as the program of their future work what they have in fact already performed.” (Glass, B., *Science and Ethical Values*. London 1966. S. 89, zit. nach: Spier, Peer review and innovation, op. cit., S. 107f.) Es gibt verschiedene Strategien, Ideen Diebstahl zu verhindern. Nach der Entdeckung eines neuen Hochtemperatur-Supraleiters bauten Paul Chu von der University of Houston und Maw-Kuen Wu von der University of Alabama in ihr zu begutachtendes Manuskript einen Fehler ein, um anderen Arbeitsgruppen, die durch „Lecks“ im Begutachtungssystem Informationen erhielten, die sie eigentlich nicht erhalten durften, keinen Vorteil zu geben. Anstelle des Elements Yttrium nannten sie das Element Ytterbium. Bei den Fahnenkorrekturen, das heißt sehr kurz vor der Veröffentlichung, änderten sie die Bezeichnung – mit der Erklärung, es wäre ein „Tippfehler“ gewesen. Offenes Mißtrauen gegenüber den Kollegen gilt offenbar als unfein. Wie berechtigt die Vorsicht war, ersieht man daraus, dass die Autoren noch während des Begutachtungsprozesses von Mitgliedern anderer Gruppen die Nachricht erhielten, mit Ytterbium würde das aber nicht funktionieren. Rustum Roy und James R. Ashburn von der Pennsylvania State University bewerten diesen Vorfall so: „Everyone except the true believers knows that it is your nearest competitors (adversaries?) who often ‘peer’ review your paper. Hence, you must protect yourself by this and other subterfuges, like proposing work you have just completed.” (Roy, R. / Ashburn, J.R., *The perils of peer review*. – In: *Nature*. 414(2001), S. 394.) Berichte, nach denen Gutachten über Projektanträge verzögert werden und die Antragsteller später erfahren müssen, dass ihre Ideen auf unbekanntem Wege zu anderen diffundiert sind, die daraus selbst einen Antrag oder eine Publikation machen, liegen sowohl aus Deutschland als auch aus den USA (teilweise in persönlichen Berichten) vor.

- ein Problem, eine Paradoxie, eine offene Flanke der gegenwärtigen Forschung innerhalb eines Bereichs bloßlegen,
- an den Begriffen, Methoden und Hypothesen einer „normalwissenschaftlichen“ Tradition feilen und polieren,
- Korrekturen an diesem Instrumentarium vornehmen, die innerhalb der Tradition, bzw. im Rahmen des Paradigmas, verbleiben,
- Theorien verschiedener Disziplinen durch Konstruktion abstrakterer Theorien verknüpfen oder vereinheitlichen und somit einen neuen, übergreifenden Gegenstandsbereich schaffen,
- ein neues Instrument vorstellen, das die Reichweite der Erfahrung vergrößert und den Vorstoß in unbekannte Dimensionen ermöglicht,
- etwas theoretisch und begrifflich Neues beginnen, das dem akzeptierten Wissen nicht widerspricht, es jedoch durch Erschließung eines neuen Wissensbereichs ergänzt und erweitert,
- eine theoretische Innovation propagieren, die geeignet sein könnte, den Rahmen eines Paradigmas, einer Tradition oder eines experimentellen Verfahrens zu sprengen,
- über die Entdeckung eines neuen, bisher für unmöglich gehaltenen Phänomens berichten,
- die bestehende Tradition von Theorie und Experiment radikal in Frage stellen und etwas fundamental Neues vorschlagen, das die Grundlagen des akzeptierten Wissens erschüttert.

Diese skalierte Typologie erhebt keinen Anspruch auf Vollständigkeit. In den Geisteswissenschaften, aber auch in den technologischen Disziplinen gibt es weitere Typen von Arbeiten. Wir lassen sie der besseren Übersichtlichkeit halber hier beiseite.

Jeder dieser Typen wissenschaftlicher Arbeit stößt auf besondere Probleme der Anerkennung und der Evaluation, aber alle Indizien deuten darauf hin, dass die Schwierigkeiten der betreffenden Forscher mit dem Peer Review System zunehmen, je weiter wir uns von oben nach unten auf der Skala bewegen. Das ist auch nicht verwunderlich, denn die Gutachter sind eine Teilmenge der Rezipienten wissenschaftlicher Ergebnisse und die Rezeption von Neuerungen ist ein Prozess, der sowohl soziale als auch kognitive Hindernisse zu überwinden hat.

Relativ gut zu bewerten sind Typen der Innovation, die am Anfang der Skala stehen. Hier kommen die konsensuellen Maßstäbe paradigmageleiteter Gemeinschaften zum Tragen. Je weiter wir auf der Skala der Originalität voranschreiten, desto unwägbarer und riskanter werden wissenschaftliche Leistungen, desto stärker wächst aber auch ihre potentielle Fruchtbarkeit und ihr möglicher Ertrag für die Zukunft.<sup>59</sup> Das Peer Review System tendiert dazu, dieses Risiko zu minimie-

ren und kleine, sichere Erträge der ungewissen Chance eines Hauptgewinns vorzuziehen. Schließlich stehen jedem „Hauptgewinn“ Dutzende von „Nieten“ gegenüber.

Dass sich die Funktionsfehler des Peer Review Systems in selektiver Weise bei der Bewertung origineller und innovativer Leistungen zeigen und konzentrieren, im übrigen aber nur in abgemildeter Form durchschlagen, ist möglicherweise einer der Gründe dafür, warum eine knappe Mehrheit der Befragten in einer Erhebung unter Psychologen sich relativ zufrieden über das Gutachterwesen des Faches äußert.<sup>60</sup>

Das Paradoxon, dass das Gutachtersystem ungeachtet seiner aktenkundigen Funktionsfehler offenbar von den meisten Wissenschaftlern akzeptiert und sogar als nützlich empfunden wird, könnte darin seine Erklärung finden, dass diese Funktionsfehler nur bei einer Minderheit voll durchschlagen. Man darf daraus

59 Außer der Achse Innovation-Tradition gibt es noch andere Dimensionen wissenschaftlichen Arbeitens, die dem Peer Review System besondere Schwierigkeiten machen. Eine davon betrifft die Generalist-Spezialist Dichotomie. Notorisch schwierig zu beurteilen ist jener Typ von Wissenschaftler, den man als „Generalisten“ bezeichnet. Dieser Typus ist nicht einheitlich. Wir finden Generalisten, die zugleich Spezialisten auf bestimmten Gebieten sind – in der Regel solche, die sich bemühen, über den Rand des eigenen Tellers zu sehen und die allorts aufgerichteten und eifrig verteidigten intellektuellen Zäune und Grenzen zu überwinden. Diese Anstrengung ist ebenso notwendig wie schwierig; sie sollte sorgfältig diskutiert und behutsam korrigiert, aber nicht vorschnell diskreditiert werden. Ein zweiter Typ von Generalist ist jener, der sich in der Konstruktion allgemeiner Theorien versucht. Das extremste, wenngleich nicht typischste Resultat eines solchen Versuchs ist eine fachübergreifende Theorie wie die Systemtheorie, die Kybernetik, die Theorie der autopoietischen, selbstorganisierenden Systeme, die Theorie der dissipativen Strukturen, die Chaostheorie und die Synergetik. Unter dieser Perspektive erscheinen Generalisten als „Spezialisten fürs Allgemeine“. Auch wenn es nicht immer zu jenen großen Theorienentwürfen kommt, hat der Generalist eine wichtige Funktion in der Wissenschaft. Es gibt zuweilen Parallelen in den Problemen, Theorienbildungen und Lösungswegen verschiedener Disziplinen, die dem Spezialisten leicht entgehen. Die Wissenschaftsgeschichte zeigt, dass Disziplinen voneinander lernen können. Zur Initiierung solcher Lernprozesse braucht die Wissenschaft, ungeachtet ihrer unaufhaltsamen Ausdifferenzierung, auch weiterhin den Generalisten, der sich unter Inkaufnahme aller sich daraus ergebenden Nachteile und Defizite darum bemüht, den Überblick zu behalten. Der Generalist kann der speziellen Forschung helfen, Analogien zu erkennen, Holzwege zu vermeiden und Anregungen für die eigene Begriffs- und Hypothesenbildung zu gewinnen. Die Institutionen der Wissenschaft müssen die Lektionen ihrer Geschichte beherzigen und aus ihnen lernen. Dazu müssten sie diese freilich besser als bisher kennen. Sie müssten ein fallbezogenes „Gedächtnis“ dafür entwickeln, wann mit welchen Mitteln und unter welchen Bedingungen und Konstellationen wissenschaftliche Erfolge erzielt wurden und ihre Entscheidungen daran orientieren – immer in dem Bewusstsein, dass niemand den Erfolg garantieren kann.

60 Krampen, G. / Montada, L., *Wissenschaftsforschung in der Psychologie*. Göttingen: Hogrefe 2002. S. 57.

nicht den Schluss ziehen, dass die Fehler des Systems tolerabel sind, weil sie von der Mehrheit wegen Nichtbetroffenheit hingegenommen werden. Vielmehr wirkt die nicht intendierte, aber in die Funktionsweise des Systems fest eingebaute und unter dem Schild der Seriosität und Solidität firmierende Beharrungstendenz als Innovationsbremse, die den Fortschritt der Wissenschaft massiv behindert.

Woher wissen wir das? Wir wüssten es aus der Wissenschaftsgeschichte, selbst wenn das Peer Review System für uns nur eine Black-Box wäre.

Die Wissenschaftsgeschichte zeigt, dass abweichende wissenschaftliche Meinungen, die heute am Peer Review System scheitern, indem sie als theoretisch abwegig, methodisch schlecht definiert oder praktisch nutzlos und somit im Sinne moderner Evaluationskriterien als „ineffektiv“ bewertet werden, morgen als richtig, wertvoll und zukunftsweisend anerkannt werden können.<sup>61</sup> Aber auch der umgekehrte Prozess lässt sich belegen: Einst hochgeschätzte Ideen, Theorien und Wissenschaftler können innerhalb weniger Jahre oder Jahrzehnte in Vergessenheit geraten. Diese andere Seite der Funktionsweise des Peer Review Systems wird zu meist unterschlagen oder vernachlässigt. Aber die Analyse der Profiteure ist genauso wichtig wie die der Verlierer. Wissenschaftliche Wertschätzung scheint eine Variable zu sein, die von der Zeit und ihren Umständen abhängt.

#### 4. *Wie kann man das Peer Review Verfahren verbessern?*

Wir hatten zu Beginn das Peer Review System mit dem Wetterbericht verglichen und gesagt, mit dem einen stehe es so ähnlich wie mit dem anderen. Das stimmt nicht ganz. Der Wetterbericht wird nämlich langsam besser – zumindest behaupten die Meteorologen das. Gilt das auch für das Gutachtersystem?

In der Literatur zum Thema werden Vorschläge zur Verbesserung des wissenschaftlichen Begutachtungssystems kontrovers diskutiert.<sup>62</sup> Unstrittig ist, dass die Güte der Entscheidungen von der Kompetenz der Gutachter und der von ihnen investierten Zeit abhängt. Unkundige Gutachter und hastig geschriebene Beurteilungen schaden der Wissenschaft. Doch soll man daraus schließen, dass man für den Peer Review Prozess nur die besten Forscher heranziehen darf und dass auch diese nur mit größter Sorgfalt begutachten dürfen? Die Konsequenz wäre, dass die besten Wissenschaftler mit der Abfassung von Gutachten überlastet wären und nicht mehr die Zeit fänden, ihre eigenen wissenschaftlichen Pläne zu ver-

61 DiTrocchio, F., *Newtons Koffer. Geniale Außenseiter, die die Wissenschaft blamierten.* Frankfurt: Campus 1998.

62 Siehe zum Beispiel: Fröhlich, G., *Anonyme Kritik. Peer Review auf dem Prüfstand der Wissenschaftsforschung.* a.a.O., S. 129 – 146; Daniel, H.-D., *Guardians of Science*, op. cit., Kap. 11; Kostoff, *Research program peer review*, op. cit., passim.

folgen. Je besser sie begutachten würden, desto größer wäre die Zufriedenheit der Auftraggeber und desto zahlreicher die künftigen Aufträge.

Obwohl diese Strategie dem Peer Review Verfahren vermutlich zu einer höheren Reliabilität verhelfen würde,<sup>63</sup> könnte dieser Vorteil den Schaden für die Wissenschaft, der durch die Bindung höchster Kompetenz entstehen würde, kaum ausgleichen. Jeder Versuch der Verbesserung des Systems muss die richtige Balance zwischen Aufwand und Ertrag finden, die den Nutzen des Gesamtsystems optimiert und zugleich die Leistungen der einzelnen Wissenschaftler gerecht beurteilt.

Um die negativen Auswirkungen von Fehlentscheidungen der Peers abzumildern, wurde von kanadischen Wissenschaftlern anstelle des gängigen „Alles-oder-Nichts“ Verfahrens ein gestuftes System der Drittmittelförderung vorgeschlagen.<sup>64</sup> In diesem System gibt es Abschlüsse für Projekte, die die Gutachter nicht völlig überzeugt haben, Zuschläge für Forscher oder Arbeitsgruppen, die bisher durchgängig hervorragende Arbeit geleistet haben, sowie gleitende Übergänge, wenn ein Antrag auf Weiterförderung abgelehnt wurde. Anfänger ohne „track of records“ erhalten einen Bonus, der ihnen die Chance gibt, sich einen guten Ruf zu machen und „intellektuelles Kapital“ anzusammeln. Es ist klar, dass dieses Verfahren im Publikationswesen nur schwer umzusetzen ist, es sei denn, man quotiert Druckseiten nach einem Muster, das hervorragend beurteilten Aufsätzen ein ungekürztes Erscheinen ermöglicht, kritisch beurteilten aber nur Raum für eine mehr oder weniger stark gekürzte Fassung zugesteht. Netzadressen können Interessenten den Weg zur ungekürzten Fassung oder zu weiterem Material weisen.

Es wurde vorgeschlagen, den Begutachtungsprozess durch Schulung der Gutachter und durch schriftliche Handreichungen, also Anleitungen für die Durchführung der Begutachtung zu verbessern. Doch wie weit kann man Gutachter überhaupt trainieren? Man kann ihnen Techniken beibringen und sie anleiten, auf bestimmte Punkte zu achten. Doch kann man sie auch Weitsicht, Realitätsinn, Mut zum Risiko und Toleranz gegenüber unkonventionellen Ideen<sup>65</sup> lehren?

Verschiedene Kritiker haben vorgeschlagen, das Peer Review System durch Veränderung von technischen Aspekten seiner Durchführung zu verbessern. Hier liegt in der Tat vieles im argen. Gerhard Fröhlich kritisiert zu Recht die „Arkanpraxis der Zeitschriftenverlage, deren Herausgeber-Referee-Begutachteten-Interaktionen nach

63 Dies ist bereits aus dem Grund zu erwarten, dass die Zahl der Gutachter relativ klein wäre und somit eine statistische Häufung von Gutachter-Paaren auftreten müsste.

64 Vgl. Kostoff, op. cit., passim. Kostoff nennt es das „bicameral system“.

65 Vgl. hierzu Fischer, K., Die Funktion der Toleranz in der Ökologie des Wissens. – In: Yousefi, H.R. / Fischer, K. (Hrsg. v.), Die Idee der Toleranz in der interkulturellen Philosophie. Nordhausen: Bautz 2003. S. 51 – 83.

dem Modell konspirativer Organisationen funktionieren“.<sup>66</sup> Diese kafkaeske Praxis vieler Zeitschriften und Drittmittelförderer, die den Anbieter eines Wissenschaftsprodukts als Bittsteller behandelt, der der Willkür eines völlig undurchsichtigen Entscheidungsverfahrens unterworfen ist und bei Ablehnung mit einem substanzlosen Formbrief abgespeist wird, der nicht auf die Gründe der Ablehnung eingeht, ist nicht hinnehmbar. Kritik ist einer der stärksten Motoren des Wissenschaftsprozesses. Aber Kritik muss überprüfbar und ihrerseits kritisierbar sein.<sup>67</sup> Und sie muss offen erfolgen, denn nur dann können alle Beteiligten aus ihr lernen.

Bisher ist es nur gnädige Kulanz der Drittmittelförderer oder Zeitschriften, wenn der Antragsteller hin und wieder (in der Regel zensierte) Stellungnahmen der Gutachter zu Gesicht bekommt. Ein Recht darauf hat er nicht. Er hat auch nicht das Recht, formell Widerspruch einzulegen und damit eine erneute Behandlung seines „Falles“ vor einer anderen Instanz – gewissermaßen einer zweiten Kammer – zu verlangen. Angesichts der Funktionsfehler des Systems und der einschneidenden Konsequenzen einer Fehlentscheidung muss das Verfahren transparenter gestaltet werden und das Recht auf ein „Berufungsverfahren“ fest verankert werden. Bei unauflösbaren Konflikten sollte der Antragsteller das Recht haben, seinen „Fall“ vor dem Tribunal der Fachöffentlichkeit – etwa auf speziell dazu eingerichteten Internetforen – diskutieren zu lassen, und zwar unter Einschluss aller am bisherigen Verfahren Beteiligten.<sup>68</sup> Vielleicht sollte es Zeitschriften geben, die der Analyse solcher Konflikte (nach dem Muster von *The Behavioral and Brain Sciences* oder *Erwägen – Wissen – Ethik*) Raum geben. Dies kann nicht für alle Ablehnungen gelten, aber die Analyse charakteristischer Fälle könnte bei Gutachtern und Antragstellern das Bewusstsein dafür schärfen, dass eine Entscheidung immer vor dem Hintergrund bestimmter Voraussetzungen getroffen wird und dass die Kriterien für eine gute wissenschaftliche Arbeit oder ein gutes wissenschaftliches Projekt aus unterschiedlichen Perspektiven ausgelegt werden können.

Über die Einzelheiten der notwendigen Veränderungen kann man sicher streiten. Ob offene Begutachtung (unter voller Nennung der Namen), Einfach-, Doppel- oder Dreifachblindverfahren, ist keine Frage des Prinzips, sondern der Praktikabilität und der Resultate. Wie viele Gutachter sollten es sein? Auch hier

66 Fröhlich, G., Anonyme Kritik. Peer Review auf dem Prüfstand der Wissenschaftsforschung, a.a.O. S. 130.

67 Zur empirisch vorfindbaren Handhabung von Kritik in der Wissenschaft, vgl.: Hartmann, H. / Dübbers, E., Kritik in der Wissenschaftspraxis. Buchbesprechungen und ihr Echo. Frankfurt: Campus 1984.

68 Da er damit auch ein persönliches Risiko eingeht, ist die Gefahr, dass das Kommunikationssystem der Wissenschaft durch eine Flut redundanter öffentlicher Debatten überlastet wird, eher gering. Vom technischen Standpunkt aus wäre dies ohnehin kein Problem.

sollte man offen sein. Wer jeweils die Mehrheitsmeinung der wissenschaftlichen Gemeinschaft zur Geltung kommen lassen will, wird eine möglichst große Zahl von Gutachtern, gewissermaßen eine „demokratische Abstimmung“, fordern;<sup>69</sup> wer ein eher elitäres Bild der Wissenschaft hat, wird einen oder zwei möglichst kompetente Forscher für ausreichend halten.<sup>70</sup>

Beide Lösungen funktionieren am besten (aber immer noch schlecht) bei der Herstellung einer Rangordnung unter dem „Gewohnten“. Wenn es um die Bewertung innovativer Forschung und neuer Ideen geht, die das kleinkalibrige Format überschreiten, sind sie strukturell überfordert.<sup>71</sup> Hier kann nur ein hinreichend vielfältiger und bunter Markt an Förderungsmöglichkeiten und Publikationsorganen Abhilfe schaffen, der auch hochgradig konsensverletzenden Ansichten eine Chance lässt.

Richtig erscheint allerdings, dass derjenige, der über die Auswahl der jeweiligen Gutachter entscheidet, eine Schlüsselposition innehat. Diese kann zum Wohle oder auch zum Schaden der Wissenschaft eingesetzt werden. Es ist essentiell, dass dies nicht den persönlichen Qualitäten des Betreffenden überlassen bleibt, sondern dass ein Verfahren installiert wird, das strukturelle Sicherungen gegen Willkür enthält. Denkbar wäre der Aufbau eines computergestützten Expertensystems, das die Kompetenzmerkmale aller potentiellen Gutachter enthält und zumindest eine Vorsortierung unter diesen nach überprüfbareren Kriterien erlaubt. Ob der Pool der potentiellen Gutachter alle überhaupt in Frage kommenden Wissenschaftler oder nur einen Teil von ihnen einschließen soll, wäre zu diskutieren. Dass es wünschenswert ist, den Kreis der Gutachter zu erweitern, scheint unter den „Benutzern“ des Systems kaum strittig zu sein. Insbesondere dem

69 Die Nachteile eines solchen Verfahrens wurden bereits in den obigen Erörterungen klargemacht. Wissenschaft funktioniert nicht demokratisch: Über Wahrheit kann man nicht abstimmen. Wissenschaftliche Durchbrüche werden immer von Minderheiten, oft von einzelnen, erzielt, die gegen den breiten Strom schwimmen und dem „Pied Piper-Effekt“ (Kostoff) nicht zum Opfer gefallen sind.

70 Allerdings erkennen auch die „kompetenten Forscher“ einen wissenschaftlichen Durchbruch oder eine erfolgversprechende innovative Idee nicht immer als solche(n).

71 Friedhelm Neidhardt formuliert es in seiner DFG-Studie sehr vorsichtig. „Unterstellt man, dass Fachgutachter innerhalb ihrer Fächer im Durchschnitt überdurchschnittlich qualifiziert sind [...], dann lässt sich annehmen, dass der Selbststeuerungszirkel der DFG für ‚kleinere Fortschritte‘ auch im ungünstigsten Fall durchaus tauglich ist. Wer sich für diesen Fall jedoch ‚Wissenschaftsrevolutionen‘, also große Durchbrüche und ‚schöpferische Zerstörung‘ der herrschenden Standards wünscht, wird freilich nicht darauf setzen können, dass Ansätze dazu – wenn es sie denn gibt – von der DFG systematisch wahrgenommen, ermutigt und unterstützt werden. Selbststeuerung sichert eher den Weiterlauf von ‚normal science‘.“ (Neidhardt, F., Selbststeuerung in der Forschungsförderung. Das Gutachterwesen der DFG, a.a.O., S. 136)

Verdacht, dass das gegenwärtige System zur Bevorzugung bestimmter Forschungsrichtungen, Paradigmen oder Methoden führt, könnte damit begegnet werden.

Wenn die Gutachtergremien einen repräsentativen Querschnitt der Disziplin bilden würden, könnte jeder Vertreter die Anträge aus seinem paradigmatischen (methodischen...) Bereich begutachten und das Problem der Benachteiligung von Alternativen zur Majoritätsmeinung wäre zumindest gemildert. Die zu verteilenden Mittel würden einfach in Proportion zur Zahl der Anhänger einer bestimmten Richtung vergeben – wobei den „Richtungsvertretern“ die Aufgabe zukäme, für die Verteilung in ihrem Bereich zu sorgen. Dies wäre eine Art „Mehrparteiensystem ohne Fünfprozenthürde“, in dem jede Partei in Proportion zu ihrer Größe an der Regierung beteiligt wird. Dies würde im politischen Bereich sicher nicht funktionieren, wäre aber ein Ansatz zur gerechteren Verteilung von Forschungsmitteln. Ohne zusätzliche Regelungen wäre das System allerdings auch schwerfällig, weil es neue Ideen nicht selektiv, sondern nur in Proportion zur Zahl ihrer jeweiligen Anhänger, fördern würde.

Wäre dieses System besser als das bisherige? Es wäre vielleicht ein wenig besser, aber einen Sprung zu einer neuen Qualität würde es nicht markieren. Warum nicht? Weil es immer hinter dem aktuellen Zustand der Disziplin herhinken müsste, niemals die ganze Komplexität ihres Bereich abbilden und das Problem der Verteilung *innerhalb* der markierten Bereiche nicht zufriedenstellend lösen würde. Dies sieht man an sogenannten monoparadigmatischen Disziplinen. In solchen Disziplinen sollte es eigentlich aus inhaltlichen Gründen keine Benachteiligungen geben. Doch es gibt sie.<sup>72</sup> Eine Analogie aus der Politik mag hilfreich sein: Es ist plausibel, dass in einem Mehrparteiensystem die Anhänger der gerade nicht regierenden Parteien ihre Interessen nicht angemessen repräsentiert sehen. Es wäre aber ein Fehlschluss anzunehmen, dass in einem Einparteiensystem alle Anhänger der Einheitspartei gleichermaßen zufriedengestellt würden. Was wir damit sagen wollen, ist, dass auch in relativ einheitlichen Disziplinen Vielfalt im Kleinen herrscht, die zu Interessenunterschieden, Interessenkonflikten und Benachteiligungen führen kann.

Wir sagten, dass dies nur vordergründig ein Problem der Auswahl der Gutachter ist. Warum vordergründig? Weil das Hauptproblem der Mittelverteilung im Wissenschaftssystem darin besteht, dass der Kreis der Anbieter (potentiellen Antragsteller) weitgehend identisch mit dem Kreis der Abnehmer (potentiellen Gutachter) ist.<sup>73</sup> Dieses Verteilungsmodell hat gewissermaßen einen Strukturfehler: es erhebt – bildhaft gesprochen – die Lobby der Böcke zu den Wahrern der Inte-

72 Ausführliche Beispiele findet man in Fischer, K., Die Funktion der Toleranz in der Ökologie des Wissens, op. cit.

ressen der Gärtner. Anders formuliert: es widerspricht dem Prinzip der Gewaltenteilung in der Wissenschaft.

Kann man den Böcken nicht die Kunst der Gärtnerei beibringen? Direkter gefragt, kann man aus *Wissensunternehmern*, die ihrer strukturellen Lage nach Interessenvertreter sind, nicht *Treuhänder* machen, die nicht nur das Wohl ihres eigenen kleinen Bereichs (ihres Spezialgebiets, Instituts, ihrer Arbeitsgruppe, ihres informellen Netzwerks, etc.), sondern das der Wissenschaft insgesamt im Blick haben? Warum führen wir nicht obligatorische Kurse in Wissenschaftsgeschichte, Wissenschaftsphilosophie und Wissenschaftsethik für alle angehenden Forscher ein, in denen (unter anderem) die funktionale Bedeutung der inhaltlichen Toleranz gegenüber von der Mehrheitsmeinung abweichenden Ideen und die fortschrittsfördernde Wirkung einer ernsthaften Auseinandersetzung mit ihnen deutlich gemacht wird? Zweifellos eine gute Sache! Hinsichtlich der Nachhaltigkeit der hier vermittelten Einsichten ist allerdings Skepsis angebracht. Unter den strukturellen Zwängen des modernen Forschungsbetriebs – Sorge um die Anschlussfinanzierung, Abhängigkeit von der Meinung der *peers*, Koppelung der Reputation an die Höhe der Drittmittel und an die Zahl der Publikationen in *peer reviewed*-Zeitschriften mit hohem *impact*-Faktor, Abhängigkeit des Ansehens und Wohlergehens des Instituts oder der Universität vom eigenen Einwerbungserfolg – droht diese Toleranz ohne tiefgreifende flankierenden Veränderungen des Systems zu einem bloßen Lippenbekenntnis zu mutieren. Es ist sogar denkbar, dass das Problem nicht auf dieser Ebene zu lösen ist: zu viele dysfunktionale Eigeneffekte, zu viele negative und positive Rückkopplungen aufgrund falscher Signale, zu viele unerwünschte Selbstbezüglichkeiten.

- 73 Diese Behauptung kann man noch verschärfen. Es lässt sich empirisch nachweisen, dass die Gruppe der realen Gutachter (Fachgutachter und Sondergutachter) der DFG im Untersuchungszeitraum nicht nur dreimal so viele Anträge auf Forschungsmittel stellte als Nichtgutachter, sondern dass sie auch wesentlich häufiger als letztere mit ihren Anträgen Erfolg hatte (Neidhardt, F. Selbststeuerung in der Forschungsförderung, op. cit., Tabelle A5, S. 141). Die Vermutung, dass die von Gutachtern vorgeschlagenen Projekte besser waren als die der anderen, ist nicht unabhängig testbar. Zwar fällt die durchschnittliche Zahl der Publikationen bei den Gutachter etwas höher aus als bei den Nichtgutachtern (a.a.O., Tabellen A7 – A9, S., 143 – 145), aber zumindest bei den Zeitschriftenaufsätzen ist es wiederum das Peer Review System, das über Annahme oder Ablehnung entscheidet. Da auch Monographien oft aus Berichten entstehen, die im Rahmen von Forschungsprojekten entstanden sind, können Publikationshäufigkeit und Projektbewilligungen nicht als unabhängige Größen angesehen werden. Es bleibt die Vermutung, dass für die unterschiedlichen Erfolgsaussichten von Gutachtern und Nichtgutachtern andere Faktoren zumindest mitverantwortlich sind. In Frage kommen zum einen „kollegiale Rücksichten“ der Gutachter aufeinander, zum anderen aber auch der Umstand, dass man als Gutachter ein Insiderwissen darüber erwirbt, wie ein erfolgreicher Antrag aussehen sollte.

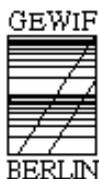
Eine bessere Lösung des Problems der Gerechtigkeit in der Verteilung der Forschungsmittel läge in der Professionalisierung des Gutachterwesens. Eine Abkopplung der Ebene der Begutachtung von der Ebene der Forschung ließe sich durch Schaffung einer Gruppe professioneller bezahlter Gutachter erreichen, die sich durch hohen Sachverstand auszeichnen sollte, keine Drittmittel einwerben darf und von ihrer Interessenlage her keine Bindungen aufweisen darf, die ihr Urteil unsachgemäß beeinflussen könnten. Natürlich sind auch professionelle Gutachter Menschen mit menschlichen Schwächen. Wie eine Partei bei Gericht einen Richter in begründeten Fällen wegen Befangenheit ablehnen kann, muss deshalb ein Antragsteller das Recht haben, in begründeten Fällen einen Gutachter abzulehnen. Über den Antrag hätte wiederum ein Gutachterausschuss zu entscheiden. Dies ist keine perfekte Lösung, aber sie ist besser als ein System, in dem ein Ausschuss von Lobbyisten darüber entscheidet, welches Verbandsmitglied welchen Anteil am Kuchen erhält.

An professionelle Gutachter sind hohe Anforderungen zu stellen. Sie müssen sich nicht nur in jener Disziplin oder jenem Spezialgebiet auskennen, deren Anträge sie zu beurteilen haben, so brauchen auch ein profundes Wissen über die Funktionsweise der Wissenschaft, das heißt: vertiefte Kenntnisse der Geschichte, Methodologie, Soziologie, Ökonomie, Philosophie und Ethik der Wissenschaften. Durch eine Ausbildung in diesen Bereichen wissen sie, wie das System arbeitet, welche Fehler man nicht machen sollte und wie man die Leistungsfähigkeit und die Dynamik des Systems erhalten und verbessern kann. Auch gut ausgebildete professionelle Gutachter können die Auswirkungen ihrer Entscheidungen auf ein prinzipiell offenes Gesamtsystem nicht vorhersehen, aber sie sind in der Lage, ihre Urteile mit Blick auf die Ökologie einer Disziplin – im Idealfall die Ökologie des Wissens insgesamt – fällen zu können.

Leider sind wissenschaftshistorische Einsichten oft nur schwer in Empfehlungen umzusetzen. Die Wissenschaft laviert, wenn sie Erfolg haben will, stets zwischen Kritik und Toleranz, Aufbau und Zerstörung, Konflikt und Kooperation, Offenheit und Abwehr, Begeisterung und Skepsis, Phantasie und Disziplin. Um den Weg zum Besseren zu finden, benötigt sie Rahmenbedingungen, die das Wechselspiel der Gegensätze ermöglichen und nicht blockieren. Sie braucht aber noch etwas anderes, nämlich Realitätssinn, Erfahrung und Kenntnis der eigenen Geschichte.

---

Gesellschaft für  
Wissenschaftsforschung



Klaus Fischer  
Heinrich Parthey (Hrsg.)

**Evaluation  
wissenschaftlicher  
Institutionen**

Wissenschaftsforschung  
Jahrbuch 2003

**Sonderdruck**

Mit Beiträgen von:

*Wolfgang Biedermann • Manfred Bonitz*

*Klaus Fischer • Siegfried Greif*

*Frank Havemann • Marina Hennig*

*Heinrich Parthey • Dagmar Simon*

*Roland Wagner-Döbler*

Wissenschaftsforschung  
Jahrbuch **2003**

---

**Evaluation wissenschaftlicher Institutionen:**

Wissenschaftsforschung Jahrbuch 2003 / Klaus  
Fischer; Heinrich Parthey (Hrsg.). Mit Beiträgen  
von Wolfgang Biedermann ... - Berlin: Gesellschaft  
für Wissenschaftsforschung 2004.

Das Werk ist in allen seinen Teilen urheberrechtlich  
geschützt.

Jede kommerzielle Verwertung ohne schriftliche  
Genehmigung des Verlages ist unzulässig. Dies gilt  
insbesondere für Vervielfältigungen, Übersetzungen,  
Mikroverfilmungen und die Einspeicherung und  
Verarbeitung in Systeme(n) der elektronischen  
Datenverarbeitung.

© Gesellschaft für Wissenschaftsforschung,  
1. Auflage 2004  
Alle Rechte vorbehalten.

Verlag:  
Gesellschaft für Wissenschaftsforschung  
c/o Prof. Dr. Walther Umstätter, Institut für  
Bibliothekswissenschaft der Humboldt-Universität zu  
Berlin, Dorotheenstr. 26, D-10099 Berlin

Druck: BOOKS on DEMAND GmbH,  
Gutenbergring, D-22848 Norderstedt

ISBN 3-934682-37-5

Preis: 15,80 €

---

# Jahrbücher Wissenschaftsforschung

Wissenschaftsforschung: Jahrbuch 1994/95.

Hrsg. v. Hubert Laitko, Heinrich Parthey u. Jutta Petersdorf. Mit Beiträgen von Siegfried Greif, Günter Hartung, Frank Havemann, Horst Kant, Hubert Laitko, Karlheinz Lüdtke, Renate Müller, Heinrich Parthey u. Manfred Wölfling. Marburg: BdWi - Verlag 1996. 306 Seiten (ISBN 3-924684-49-6) 20,00 €

Wissenschaftsforschung: Jahrbuch 1996/97.

Hrsg. v. Siegfried Greif, Hubert Laitko u. Heinrich Parthey. Mit Beiträgen von Siegfried Greif, Christoph Grenzmann, Claudia Hermann, Gunter Kayser, Karlheinz Lüdtke, Werner Meske, Heinrich Parthey, Roland Wagner-Döbler, Manfred Wölfling u. Regine Zott. Marburg: BdWi - Verlag 1998. 254 Seiten (ISBN 3-924684-85-5) vergriffen

Wissenschaft und Digitale Bibliothek: Wissenschaftsforschung Jahrbuch 1998.

Hrsg. v. Klaus Fuchs-Kittowski, Hubert Laitko, Heinrich Parthey u. Walther Umstätter. Mit Beiträgen von Manfred Bonitz, Klaus Fuchs-Kittowski, Siegfried Greif, Frank Havemann, Horst Kant, Hubert Laitko, Karlheinz Lüdtke, Heinrich Parthey, Wolfgang Stock, Walther Umstätter, Roland Wagner-Döbler, Petra Werner u. Regine Zott. Berlin: GeWif 2000. 368 Seiten. (ISBN 3-934682-30-8) 19,43 €

Wissenschaft und Innovation: Wissenschaftsforschung Jahrbuch 1999.

Hrsg. v. Siegfried Greif u. Manfred Wölfling. Mit Beiträgen von Siegfried Greif, Christoph Grenzmann, Hans-Eduard Hauser, Frank Havemann, Gunter Kayser, Andrea Scharnhorst, Roland Wagner-Döbler, Manfred Wölfling u. Janos Wolf. Berlin: GeWif 2003. 227 Seiten. (ISBN 3-934682-33-2) 13,00 €

Organisationsinformatik und Digitale Bibliothek in der Wissenschaft: Wissenschaftsforschung Jahrbuch 2000.

Hrsg. v. Klaus Fuchs-Kittowski, Heinrich Parthey, Walther Umstätter u. Roland Wagner-Döbler. Mit Beiträgen von Manfred Bonitz, Christian Dame, Klaus Fuchs-Kittowski, Frank Havemann, Heinrich Parthey, Andrea Scharnhorst, Walther Umstätter u. Roland Wagner-Döbler. Berlin: GeWif 2001. 239 Seiten. (ISBN 3-934682-34-0) 14,00 €

Wissenschaft und Innovation: Wissenschaftsforschung Jahrbuch 2001.

Hrsg. v. Heinrich Parthey u. Günter Spur. Mit Beiträgen von Wolfgang Biedermann, Manfred Bonitz, Werner Ebeling, Klaus Fuchs-Kittowski, Siegfried Greif, Christoph Grenzmann, Horst Kant, Matthias Kölbl, Rüdiger Marquardt, Heinrich Parthey, Andrea Scharnhorst, Tankred Schewe, Günter Spur u. Walther Umstätter. Berlin: GeWiF 2002. 231 Seiten (ISBN 3-934682-35-9) 15,80 €

Wissenschaftliche Zeitschrift und Digitale Bibliothek: Wissenschaftsforschung Jahrbuch 2002.

Hrsg. v. Heinrich Parthey u. Walther Umstätter. Mit Beiträgen von Manfred Bonitz, Horst Kant, Alice Keller, Matthias Kölbl, Heinrich Parthey, Diann Rusch-Feja, Andrea Scharnhorst, Uta Siebeky, Walther Umstätter u. Regine Zott. Berlin: GeWiF 2003. 222 Seiten (ISBN 3-934682-36-7) 15,80 €