Peer Review Verfahren auf dem Prüfstand

Zum Soziologiedefizit der Wissenschaftsevaluation

Peer Review Research – Reviewed

Sociological Shortcomings of Academic Evaluation

Stefan Hirschauer

Institut für Soziologie, Universität München, Konradstraße 6, D-80801 München

Zusammenfassung: Der Aufsatz bietet einen Überblick über die Hauptfelder der Peer Review Forschung, also jenes Segments der Wissenschaftsforschung, das sich mit dem zentralen Evaluationsverfahren wissenschaftlicher Praxis befasst. Er kommt zu einem kritischen Befund: Die Peer Review Forschung ist auf verquere Weise in eben die Evaluationspraxis verstrickt, die sie doch professionell beobachten müsste. Der Grund ist ein Mangel an soziologischer Durchdringung des Gegenstands. Der Aufsatz plädiert daher für eine theoretische Neuorientierung von Personen auf soziale Prozesse, von Reliabilitätsmessung auf Dissensanerkennung, von Kognition auf Sprech- und Schreibpraxis sowie von Publikationszählungen auf Kommunikationsforschung. Denn der Peer Review ist kein wissenschaftliches Messverfahren für die Güte von Publikationen, sondern eine soziale Einrichtung zur Kalibrierung der Lesezeit einer Disziplin.

1. Einleitung

Fast jeder Aspekt wissenschaftlicher Kommunikation ist durch Evaluationen bestimmt. Neben informellen Evaluationen – etwa in Betreuungsverhältnissen, durch Rezensionen oder Diskussionsbeiträge auf Tagungen – finden sich eine Reihe formeller Evaluationsverfahren. Sie lassen sich nach unterschiedlichen Aspekten differenzieren: nach den bewerteten Objekten (Personen, Organisationen, kommunikative Angebote), den verwendeten Methoden (mathematische versus qualitative Evaluation), den sozialen Grenzziehungen zwischen den Beteiligten (interne versus externe Evaluationen) sowie nach den ins Spiel gebrachten Hierarchien: von solchen des Bildungssystems (wie bei Diplom-, Promotions-, und Habilitationsprüfungen sowie studentischen Lehrevaluationen), über solche der Forschungsförderung (wie bei Auswahlverfahren für Stipendien und Expertisen, bei Projektanträgen und Institutsevaluationen) bis zu den weichen Rollendifferenzierungen von Berufungsverfahren und Manuskriptbegutachtungen.

In den meisten dieser Verfahren wird auf soziale Muster zurückgegriffen, die in einem ‚Prototyp' der Forschungsevaluation entwickelt wurden: der Beurteilung von Manuskripten durch den Peer Review von Fachzeitschriften. Diese qualitative Evaluation von Forschungsarbeiten durch *fellow scientists* wird gemeinhin als ein Kernstück wissenschaftlicher Kommunikation betrachtet, das ihren ‚organisierten Skeptizismus' institutionalisiert und gute von schlechter Forschung unterscheidet: „‚Peer-Review' steht für die Begutachtung und Bewertung … wissenschaftlicher Wissensbehauptungen durch die dazu allein kompetenten Kollegen (‚peers')" (Weingart 2001: 284f.).

Der Peer Review-Prozess umfasst eine klar umgrenzte Phase der ‚Gesamt-Biographie' eines Manuskripts: die zwischen der Einreichung bei einer Zeitschrift und der Drucklegung (vgl. Abb. 1).

In dieser Phase werden Manuskripte vorselektiert, Gutachter ausgewählt, Texte gelesen, Stellungnahmen verfasst, Herausgeberentscheidungen getroffen, Autorenbriefe aufgesetzt sowie Überarbeitungen vorgenommen und kontrolliert.[1] Vor dem Peer

[1] Die Abbildung skizziert ein elaboriertes Verfahren, wie es etwa auch die Zeitschrift für Soziologie unterhält. Dabei ist zu bedenken, dass „*peer reviewed journal*" ein Label geworden ist, mit dem sich heute die meisten Fachzeitschriften schmücken, hinter dem sich aber recht unterschiedliche Prozesse verbergen: von der Gestaltung der Personalrekrutierung (Herausgeberwahl und Gutachterauswahl) über den Grad der Formalisierung (Schriftlichkeit, fallweises oder standardisiertes Verfahren, Gutachterinstruktionen) und die redaktionelle Handhabung von Manuskripten (Anonymisierung, Gutachtenübermittlung, Verfahrensalternativen) bis zur Staffelung von ‚Instanzen' (redaktionelle Vorauswahl, unabhängige Gutachter, Alleinherausgeber oder Gremien, Beschwerdeverfahren usw.). Eine verglei-

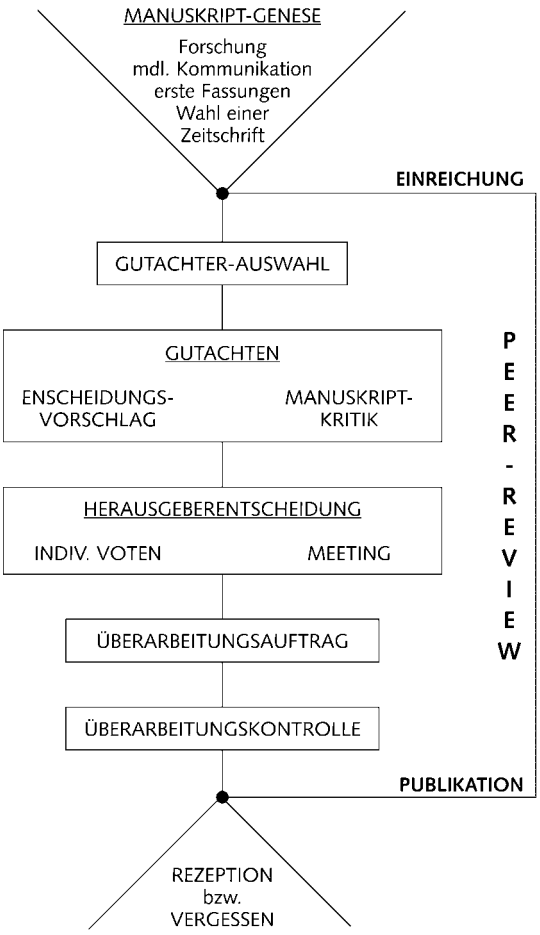

Abb. 1 Biographie einer Zeitschriftenpublikation

Review finden sich Forschungsprozesse, Vorträge und mündlicher Austausch sowie erste Fassungen eines Papiers, danach die Rezeption bzw. das Vergessen. Die „Elementarakte" (Stichweh 1994) wissenschaftlicher Kommunikation – die Publikation und die Zitation – bezeichnen also nur einen kleinen Ausschnitt der Genese kommunikativer Produkte in der Wissenschaft.

Über den Peer Review existiert seit den 70er Jahren eine rapide wachsende Forschung. Zahlreiche Disziplinen haben eine kaum mehr überschaubare Masse empirischen Wissens generiert: vor allem die Medizin und Psychologie, aber auch die Soziologie, Politikwissenschaft, Geographie, und verschiedene Naturwissenschaften. Seit der ‚klassischen' Studie von Zuckerman und Merton (1971) und einer Reihe von ihnen folgenden Monografien (Cole/Rubin/Cole 1978, Lindsey 1978, Cole/Cole 1981, Lock 1985, Neidhardt 1988, Chubin/Hackett 1990, Daniel 1993 u. a.) ist der Peer Review immer wieder Gegenstand internationaler Diskussionen gewesen. Allein in den 90er Jahren gab es vier internationale Konferenzen zum Peer Review in der Medizin, die das *Journal of the American Medical Association* (JAMA) organisierte sowie eine von Steve Fuller initiierte Internetkonferenz (Cyberconference 1999). Ferner veranstaltete die Debattenzeitschrift *Behavioral and Brain Sciences* zwei interdisziplinäre Diskussionen, die bis heute als ‚Meilensteine' der Literatur gelten können (Harnad 1982, Cicchetti 1991).[2]

Diese wissenschaftliche Aufmerksamkeit für den Peer Review ist wenig überraschend. Die Forschung hat eine doppelte Relevanz. Zum einen verspricht sie eine Selbstaufklärung der Wissenschaft über die ihr eigenen sozialen Formen der Qualitätssicherung. Zum anderen hat sie auch einen hohen Orientierungswert für Versuche der politischen Forschungssteuerung: Vor der Nutzung von ‚Peers' für die Forschungssteuerung und auch vor der Entwicklung anderer Evaluationsinstrumente für Professuren, Institute oder Forschungsfelder empfiehlt es sich, jenes Verfahren inwendig zu kennen, das als exemplarischer Fall aller Wissenschaftsevaluation gelten kann. Solange man Evaluationsinstrumente entwickelt, ohne die Mechanismen der Selbstevaluation zu begreifen, mit denen sich Wissenschaft zu einer professionellen Praxis macht, wird man das Differenzierungsvermögen jener Selbstevaluation chronisch unterschreiten.

Auf der anderen Seite weist die Peer Review Forschung gemessen an ihrer Relevanz zwei große Schwächen auf. Die erste besteht in offenkundigen empirischen Forschungslücken: Im Hinblick auf den gesamten Ablauf des Begutachtungsverfahrens gibt es eine Konzentration von empirischen Studien auf wenige ausgewählte Aspekte und eine Vernachlässigung anderer. Extrem rar sind Studien zur Aufsatz-Entstehung, zu Einreichungs- und Überarbeitungsstrategien von Autoren (Knorr 1984, Myers 1990). Ebenso schwach beforscht sind die redaktionellen ‚Ränder' des Prozesses: Sporadische Studien existieren zum Erfolg von Anonymisierungspraktiken (Rosenblatt/Kirk 1980, Ceci/Peters 1984,

chende Betrachtung und Typenbildung fehlt bislang in der Forschung.

[2] Für deskriptive Darstellungen der empirischen Resultate der Peer Review Forschung vgl. etwa Armstrong 1997, Campanario 1998a und b.

McNutt u.a. 1990, Blank 1991), kaum untersucht wurden die Gutachterauswahl (Bös 1998), die Autorenbriefe, die Entscheidungen begründen, Gutachten zusammenfassen, Widersprüche entscheiden und Überarbeitungsrichtungen nahelegen (Bonjean/Hullum 1978) sowie die Abwicklung und der Erfolg von Autorenbeschwerden (Simon/Bakanic/McPhail 1986). Über die Entscheidungsverläufe in Herausgebergremien gibt es noch gar keine empirischen Studien im eigentlichen Sinne, sondern nur Erfahrungsberichte (z.B. Roediger 1987; Perlman/Dean 1987), was auf ein Zugangsproblem hinweisen dürfte.[3] Die überwältigende Zahl von Studien konzentriert sich stattdessen auf die Publikationsempfehlungen und die Manuskript-Besprechungen der *Fachgutachter.*

Die zweite Schwäche ist theoretischer Art und liegt in den methodischen und epistemologischen Prämissen der Forschung. Nicht nur in psychologischen oder medizinischen Studien, auch in sozialwissenschaftlichen Untersuchungen findet sich einer folgenreicher Mangel an soziologischer Durchdringung des Gegenstandes. Das hat vor allem den Grund, dass die Peer Review Forschung auf komplizierte Weise in ihren Gegenstand verstrickt ist: Es handelt sich halt um eine wissenschaftliche (aber tendentiell evaluative) Forschung über wissenschaftliche Verfahren der Wissenschaftsevaluation. In dieser Forschung werden zwangsläufig mit bestimmten Methoden und Theorien auch normative Standards von ‚Wissenschaftlichkeit' eingesetzt. Man hat es also mit einem hochgradig selbstbezüglichen Geschäft zu tun, zu dessen Bewältigung das Gros der Peer Review Forschung konzeptuell unterausgestattet ist.

Dieser Aufsatz verfolgt in dieser Lage nur einen begrenzten Zweck. Er bietet zum einen einen Überblick über das interdisziplinäre Feld der Peer Review Forschung, der sich auf den Austausch von Argumenten zum Gegenstand konzentriert (2.), zum anderen identifiziert er vier konzeptuelle Schwächen dieser Forschung (3.), aus denen sich methodische Hinweise und theoretische Vermutungen für zukünftige Untersuchungen ableiten lassen.

2. Der Stand der Peer Review Forschung

Die Peer Review Forschung hat, wie gesagt, bislang weit überwiegend den Beitrag von Fachgutachtern zu den Verfahren untersucht. Dies geschah vor allem mit vier Fragestellungen. Sie richten sich auf die Vorurteilsstrukturen von Gutachtern (2.1), den Grad der Übereinstimmung ihres Urteils (2.2), die von ihnen verwendeten Gütekriterien (2.3) und die Vorhersagevalidität ihrer Empfehlungen (2.4).

2.1 Experimente zum Bias von Gutachtern

Schon die Studie von Zuckerman und Merton verfolgte vor einem klassentheoretischen Hintergrund die Hypothese eines Statusbias von Gutachtern. (Sie fand statt einer klaren Bestätigung eher einen negativen Effekt des Autoren*alters* auf die Publikationschancen von Manuskripten). Wie schon Crane (1967) vor ihnen und Cole/Cole (1978) nach ihnen arbeiteten Zuckerman/Merton dabei mit einfachen Korrelationen von Publikationsentscheidungen mit sozialen Merkmalen von Autoren und Gutachtern. Die Ergebnisse waren aus heutiger Sicht irritierend beruhigend und setzten sich u.a. dem Einwand aus, dass eine Suche nach institutionellem Partikularismus (‚old boyism', Rang(un)gleichheit, geographische Ansiedelung) die Organisation des Wissens selbst, vor allem den Bias zwischen unterschiedlichen Paradigmen ausblende (Travis/Collins 1991).

Zwei Nachfolgestudien ergänzten auch die Frage nach Personenmerkmalen durch eine Kodierung von Manuskripten nach Schulenzugehörigkeit bzw. Forschungsgebiet.[4] Sahner (1982) verglich (am Fall der Zeitschrift für Soziologie) mit einer Input-output-Analyse die Merkmale von eingereichten und publizierten Manuskripten und stellte differenzielle

[3] Tatsächlich weiß man empirisch genaueres über die Entscheidungen der Redakteure von Zeitungen (Clayman/Reisner 1998) und der Verleger von Büchern (Powell 1985) als über die Herausgeberkommunikation und das redaktionelle Geschehen in Fachzeitschriften. Wenn Peer Review Studien einmal auch Daten über diese Phase erheben (etwa Sahner 1982, Bakanic/McPhail/Simon 1987) so sind diese extrem hoch aggregiert: Eine Pfadanalyse etwa sagt nicht viel über tatsächliche ‚Pfade', d.h. Entscheidungsverläufe in Einzelfällen. Das Zugangsproblem wurde schon 1977 von Mahoney beklagt. Er berichtete von hohen emotionalen Widerständen gegen seine Studie, Drohungen von forschungsethischen Prozessen und Stellenverlust: „(an) emotional intensity that surrounds research on the peer-review system" (1977: 221). In der Tat wird gerade in der Peer Review Forschung eine gewisse empirische Wasserscheu gerne forschungsethisch bemäntelt. Die Konsequenz ist eine unberechenbare Gemengelage aus anekdotischem Insiderwissen und Verschwörungstheorien. Für analytisch-empirisches Wissen ist der operative Kern von Peer Review Prozessen noch weitgehend eine ‚black box' geblieben (Sonnert 1995).

[4] Kaum zufällig galten beide Studien soziologischen Zeitschriften.

Publikationschancen von Manuskript-Typen fest, die er einer ‚Blattlinie' der Herausgeber zuschrieb. Bakanic/McPhail/Simon (1987) zeigten mit einer Pfadanalyse Effekte der Themenfelder von Manuskripten auf ihre Publikationschance im *American Sociological Review*. Solche Studien erlauben einige allgemeine Aussagen über die Selektivität von Zeitschriften, aber diese Aussagen bleiben i.d.R. ambivalent zurechenbar: auf die personelle Zusammensetzung von Reviewprozessen oder auf die Qualität von Manuskripten (oder Autoren) einer bestimmten Kategorie.[5]

Beweiskräftiger schienen Untersuchungsdesigns, die eben jene „drakonischen Experimente" veranstalteten, die Zuckerman/Merton noch aus forschungsethischen Gründen zurückgewiesen hatten. Sie arbeiten im Prinzip mit einem Vergleich der Urteile über in bestimmten Hinsichten fingierte Manuskripte mit einer Kontrollgruppe von Urteilen über Originalmanuskripte. Solche Analysen konnten etwa zeigen, dass die Publikationschancen von Manuskripten in folgenden Fällen sinken: wenn es sich nur um Replikationsstudien handelt (s. Campanario 1998a: 203ff.), wenn empirische Ergebnisse nicht signifikant sind (s. Bornstein 1991), wenn sie dominanten paradigmatischen Orientierungen widersprechen (Mahoney 1977, Epstein 1990), und wenn sie – in Verfahren ohne Anonymisierung – von unbekannten Autoren namenloser Institutionen eingereicht werden.

Dieser letzte Bias wurde besonders eindrucksvoll in einer quasi-experimentellen Studie von Douglas Peters und Stephen Ceci (1982) zu psychologischen Fachzeitschriften demonstriert, deren Diskussion zu den erwähnten ‚Meilensteinen' der Peer Review Forschung gehört. Schon die Publikation ihres Aufsatzes ist insofern interessant, als er zum einen zuvor von anderen prestigereichen Zeitschriften (*Nature* und *American Psychologist*) abgelehnt wurde, zum anderen schließlich in einer Zeitschrift (den *Behavioral and Brain Sciences*) erschien, die ‚*open peer commentary*' praktiziert, d.h. ihre Aufsätze nach dem internen Peer Review auch einer öffentlichen Debatte aussetzt. An dieser nahmen etwa 60 Autoren teil, darunter 23 aus der Psychologie, 20 aus Naturwissenschaften und Medizin sowie 16 aus Sozial- und Verhaltenswissenschaften. Die Diskussion selbst erwies sich als ausgezeichnetes Dokument für die oft kritisierte ‚Reliabilitätsschwäche' wissenschaftlicher Meinungsäußerungen: Es gab ebenso begeisterte wie vernichtende Kommentare.

Peters/Ceci wählten aus 12 angesehenen psychologischen Zeitschriften (mit hohen Rejektions- und Zitationsraten) je einen Artikel der letzten 2-3 Jahre aus, und veränderten Autoren- und Institutionennamen vom Hochreputierten ins Namenlose.[6] Außerdem nahmen sie einige kosmetische Veränderungen am Titel, am Abstract und an den Einleitungsabschnitten vor. Dann reichten sie den ansonsten unveränderten Artikel bei der gleichen Zeitschrift ein, die ihn schon publiziert hatte. Das erstaunliche Ergebnis: Nur 3 der 12 Manuskripte wurden als bereits publiziert erkannt, und 8 der 9 übrigen wurden nun *abgelehnt*.

Die Diskutanten warfen neben den methodischen Schwächen der Studie vor allem zwei Fragen auf. Die erste war: Warum wurden die Manuskripte nicht als bereits publiziert, als ‚Plagiate', erkannt? Es wurden vor allem drei Gründe vorgeschlagen. Zum ersten könnte die Begutachtungspraxis der untersuchten Zeitschriften dafür sorgen, dass Herausgeber wie Gutachter die Artikel nicht kannten. Da die Herausgeber in 7 von 9 Fällen identisch waren, muss man annehmen, dass sie über die Manuskripte allein aufgrund der Gutachten, also ohne eigene Lektüre entschieden (Goodstein).[7] Dafür spricht, dass in Zeitschriften mit hohen Rejektionsraten u.U. einige Hundert Manuskripte pro Jahr zu entscheiden sind, was es nahelegt, die Urteile zweier Gutachter als hinreichende Vorauswahl von überhaupt ‚Lesenswertem' einzusetzen, *wenn* diese in der Ablehnung übereinstimmen. Dass aber auch fast allen Gutachtern die Plagiate nicht auffielen, macht es zunächst eher unwahrscheinlich, dass es sich um dieselben Personen wie bei der Begutachtung der Originalpublikation handelte. Ferner könnten sie wie bei vielen Zeitschriften so ausgewählt worden sein, dass sie eben nicht exakt im Feld des Manuskripts, sondern nur in seiner Nähe arbeiten und ein gutes Urteil aus der Halbdistanz versprechen, aus der Plagiate nur selten bemerkt werden (Scarr).

Der zweite Grund generalisiert das letzte Argument mit Hinweis auf den kleinen Leserkreis, den Fachaufsätze im allgemeinen haben. Mahoney (1987)

[5] Ebenso argumentierten denn auch die damaligen Herausgeber der ZfS in ihrem auf Sahners Artikel reagierenden Editorial von 1982.

[6] Die Harvard University wurde etwa zum „Northern Plains Center for Human Understanding".

[7] Namensnennungen ohne Jahresangabe bezeichnen hier wie im Folgenden die Autoren von Diskussionsbeiträgen in Peters/Ceci bzw. Harnad 1982. Ebenso verfahre ich in einem späteren Abschnitt mit den Diskutanten um Cicchetti 1991.

schätzt wie schon Garvey/Griffith (1971), dass ein durchschnittlicher Aufsatz etwa 1 % der Leser einer Zeitschrift findet. Mehr als die Hälfte aller Artikel wird niemals zitiert (Garfield 1989: 7). Die durchschnittliche Zitationsrate sozialwissenschaftlicher Artikel ein Jahrzehnt nach Erscheinen liegt bei 1,4 Zitaten (Garfield 1979). Auch die von Peters/Ceci ausgewählten Zeitschriften konnten schon als ‚vielzitiert' gelten, weil ihre Artikel nach ihrem Erscheinen durchschnittlich 1,15 Zitate pro Jahr erhielten (die von Peters/Ceci ausgewählten Artikel 1,5 Zitate). Der hohe Grad von disziplinärer Spezialisierung und die Zersplitterung von Forschungsfeldern macht die Lektüre, die Wissenschaftler zur Fortsetzung ihrer eigenen Arbeit aufrechterhalten können, extrem ausschnitthaft. Es ist daher gut möglich, dass die Gutachter die Originalmanuskripte nie gelesen hatten.

Aber selbst wenn sie sie lasen, gibt es noch einen dritten Grund für die Verkennung ihres Plagiatcharakters: die geringe Individualität von Fachaufsätzen. Gerade in Peters und Cecis Feld, der Psychologie, findet sich eine hohe Standardisierung von Methoden und Darstellungsweisen. Wenn Papiere sich aber ohnehin stark ähneln, kann man exakte Kopien leicht übersehen, selbst wenn Artikel doch schon einmal gelesen wurden. Einige Diskutanten zeigten sich denn auch wenig überrascht, dass die Plagiate so selten aufflogen. Wenn einerseits Autoren massenhaft redundante (nur marginal veränderte) Aufsätze publizieren und andererseits nur ein Bruchteil der Leser einer Zeitschrift einen Aufsatz liest, dann ist die ‚Entdeckungsrate' in der Studie von Peters/Ceci recht hoch (D. Beaver). Douglas Eckberg verweist darauf, dass es in der Psychologie ca. 180 englischsprachige Peer Review Zeitschriften mit Tausenden jährlicher Publikationen gibt. Die meisten sind ‚*normal science paper*', die ‚ganz ordentlich', aber keineswegs bahnbrechend innovativ sind. Ihr Erinnerungswert sei daher gering, auch wenn sie für spezielle Forschungsinteressen einen kleinen Beitrag gemacht haben mögen. Auch Katherine Nelson meint, die meisten ‚publizierbaren' Manuskripte würden schnell von Herausgebern, Gutachtern und Lesern vergessen, einfach weil sie schlicht vergessenswürdig sind. Dauerhaft wichtige Arbeiten sind sehr rar in einem Forschungsfeld. Und die Arbeiten in psychologischen Labors sind nur für eine sehr kleine Zahl von Spezialisten relevant, die für einen begrenzten Zeitraum an ähnlichen Fragen arbeiten.

Das zweite zentrale Thema der Diskussion um Peters/Cecis Studie war die Erklärung der absurd erscheinenden Ablehnung vormals akzeptierter Manuskripte. Die meisten Diskutanten zeigten sich von der Hypothese eines Statusbias überzeugt und legten entsprechende Anonymisierungsvorkehrungen nahe. Es gab aber auch eine Reihe von Einwänden gegen die Hypothese sowie zwei konkurrierende Erklärungsversuche.

Zunächst lässt die Studie aufgrund ihres Designs offen, ob sie nun einen Bias zugunsten von prestigereichen Institutionen und Autoren oder einen zuungunsten von prestigearmen aufdeckte, ja es könnte, wie viele Diskutanten anmerken, aufgrund der Namenwahl sogar ein (wenig überraschender) Bias gegen nicht-akademische Institutionen sein. Wie aber ist dieser zu bewerten? Vor allem die Physiker unter den Diskutanten zeigten sich eher ungerührt von Peters und Cecis Bestätigung des Mertonschen ‚Matthäus-Effekts' (1968/1985), der aus der Perspektive einer ‚Aufdeckung von Ungleichheiten' formuliert wurde. Wissenschaft sei nunmal keine demokratische Angelegenheit, die Orientierung an Reputation reduziere Komplexität, und dies, wie Daniel Perlman an Zitationshäufigkeiten demonstriert, recht erfolgreich.

Ferner wurde diskutiert, auf welche Weise der Statusbias wirken könnte. Die Anlage der Studie könnte auf zwei Weisen eine spezifische Reaktivität ausgelöst haben. Zum einen im Hinblick auf die Votenstile der ausgewählten Gutachter: Daryl Chubin gibt zu bedenken, dass der Status von Autoren in den Augen von Herausgebern entscheidend sein kann, um ihre ‚Peers' zu bestimmen. Die Artikel könnten daher auch abgelehnt worden sein, weil die Herausgeber entsprechend der institutionellen Herkunft weniger reputierte (vor allem: jüngere) Gutachter auswählten, deren Urteil nach den Ergebnissen verschiedener Studien (Snizek/Fuhrman 1979, Amabile/Glazebrook 1982) i.d.R. schärfer ausfällt als das von ‚*Seniors*'. Zum anderen weisen Manwell/Baker darauf hin, dass es sich bei den (wegen ihrer gehobenen Zitationsraten) ausgewählten Aufsätzen um ‚*high-risk*'-Papiere handeln könnte, deren Claims (anders als die meisten Aufsätze) starke Reaktionen bei Gutachtern auslösen, weil sie typischerweise Innovativität mit handwerklichen Mängeln kombinieren. Solche Mängel dürften Herausgeber wie Gutachter aber eher reputierten Autoren ‚durchgehen' lassen.

Schließlich wurde die Überzeugungskraft der Status-Hypothese bezweifelt: Aufgrund der kleinen Fallzahl müsse unentschieden bleiben, inwieweit es sich überhaupt um einen systematischen Bias handelt oder ob der ‚Zufall', d. h. eine Vielzahl weiterer denkbarer Faktoren, zu dem Ergebnis führte, darunter die Rejektionsquote der Zeitschriften, die

uneindeutige, nämlich weder herausragende noch indiskutable Qualität der meisten Manuskripte (die man drucken kann, wenn es der Platz erlaubt) und Fehlurteile oder einfache Meinungsverschiedenheit der alten und neuen Gutachter. Domenic Cicchetti weist auf den Fall hin, dass auch ein einmütig akzeptiertes Manuskript nach seiner Publikation entschiedene Kritik evozieren kann, deren frühzeitige Mobilisierung die Publikation auch hätte verhindern können. Ganz in diesem Sinne bekennt Robert Rosenthal, dass er den Artikel von Peters/Ceci, sollte er in einer zukünftigen Peer Review Studie einmal für einen Reliabilitätstest wiedereingereicht werden, mit Sicherheit nicht zur Publikation akzeptieren würde. Insofern würde er (mit seinen kritischen Einwänden zur Methode) die von den Autoren demonstrierte ‚Reliabilitätsschwäche' des Peer Review bestätigen. Stanley Presser meint, solche Faktoren reichten völlig aus, um anzunehmen, dass die Manuskripte u. U. ebenso abgelehnt worden wären, wenn Peters/Ceci sie nicht verfälscht hätten.

Über diese Einwände hinaus wurden aber auch zwei konkurrierende Hypothesen zur Ablehnung der Manuskripte formuliert. Die erste besteht in dem Argument, dass die Manuskripte in den zwei bis drei Jahren nach ihrem Erscheinen eine für ihre Publikation entscheidende Eigenschaft verloren, die auch Peters und Cecis experimentelles Design nicht ‚konstant' halten konnte: ihre Neuheit. Vor dem Hintergrund der schwachen Erinnerbarkeit von Aufsätzen könnten sie als ‚veraltet' gelten, ohne als Plagiat erkannt zu werden. Donald Beaver meint, auch wenn Gutachter einen Artikel nie gelesen haben, können sie seinen Inhalt durch Vorträge, Preprints oder persönliche Mitteilungen kennen, und dies auch schon lange vor der Publikation. Sie können einen Aufsatz (nach Lektüre von ein paar Hundert anderen) daher implizit als ‚veraltet' wahrnehmen, auch ohne ihn identifizieren zu können.

Die zweite Konkurrenzhypothese zum Statusbias bestreitet eine zweite implizite Konstanzannahme des experimentellen Designs. Was Peters und Ceci als (heimliche) *Wieder*einreichung konzipierten, musste von Gutachtern, die dies nicht erkannten (und auch nicht erkennen sollten!), als *Erst*einreichung verstanden werden. Solche Einreichungen werden aber mit einer strengeren Haltung begutachtet als Manuskripte, die als ‚Überarbeitungen' auftreten. Darüber hinaus konnten Peters/Ceci mangels Einblick in die ersten (‚originalen') Reviewprozesse nicht ausschließen, dass die Originalbeiträge das Endresultat mehrfacher Überarbeitungen und vielleicht sogar Einreichungen waren. Während die vermeintlichen Wiederholungen also als Ersteinreichungen erschienen, könnten die vermeintlichen Ersteinreichungen de facto Wiederholungen gewesen sein. Die hohe Rejektionsrate der verfälschten Manuskripte kann also schlicht auf eine ungleiche Behandlung von ersten und wiederholten Einreichungen durch Gutachter verweisen. Bei Ersteinreichungen kann eine Zeitschrift mehr von Autoren fordern, und dokumentiert ohne Zögern, was sie von ihnen und von sich selbst erwartet (konkret: was Gutachter annehmen, was Herausgeber erwarten). Nehmen Autoren diese Hürde und reichen ein zweites Mal ein, so werden ihre Manuskripte für die Anforderungen bestimmter gutachtender Leser zugeschnitten und geschliffen (D. Rubin, K. Nelson). In den Zweiteinreichungen können Gutachter durch ihre Überarbeitungsauflagen also zahlreiche Spuren hinterlassen haben, evtl. sogar zu latenten Koautoren geworden sein.[8] Es ist dann überhaupt nicht überraschend, dass andere Gutachter andere Ansprüche haben und zu anderen Urteilen kommen. Eben deshalb zählt es bei den meisten Zeitschriften zu den stillen Übereinkünften zwischen Herausgebern und Autoren, dass der Prozess der ‚Optimierung' von Manuskripten irgendwann abgebrochen wird, um Publikationen nicht auf unfaire Weise – nämlich durch Überziehen der temporären Rollendifferenzierungen zwischen den Peers – zu erschweren und zu verzögern.

2.2 Messungen der Reliabilität von Urteilen

Neben den Versuchen zum Nachweis von Vorurteilen fragt eine zweite große Gruppe von Forschungen über den Peer Review nach dem Grad der Übereinstimmung zwischen den Urteilen von Gutachtern. Solche Studien treten entweder in der Tradition von Zuckerman/Merton als diziplinvergleichende Untersuchungen von Paradigmenkonsens oder (vor allem in der Psychologie) als Reliabilitätsmessungen auf. Entsprechende Untersuchungen stammen u. a. von McCartney (1973), Cole/Rubin/Cole (1978), Lindsey (1978), Cicchetti/Eron (1979), Cole/Cole/Simon (1981), Root (1987), Hargens/Herting (1990a), Daniel (1993), Junge (1993) u. a.

Die bislang umfassendste Studie zur Übereinstimmung von Gutachterurteilen stammt von Domenic Cicchetti (1991), der die Reliabilität des Peer Review in verschiedenen Disziplinen verglich und als disziplinübergreifend schwach diagnostizierte. Ein weiteres Ergebnis von Cicchetti ist, dass der Rejek-

[8] Diesen Punkt bezweifelt einer der anonymen Gutachter dieses Aufsatzes. Dem anderen sei (u. a.) für den Vorschlag zu seinem Titel gedankt.

tionskonsens in allen untersuchten Disziplinen deutlich höher als der Akzeptanzkonsens ist. Dies gilt vor allem für die Begutachtung von Projektanträgen und für die von Manuskripten allgemeindisziplinärer Zeitschriften.

Die Diskussion des Aufsatzes, die wie die von Peters und Cecis Studie in den *Behavioral and Brain Sciences* publiziert wurde, widmete sich zunächst auf breiter Front messtechnischen Fragen. Diese waren schon vor Cicchettis Studie kontrovers (s. etwa Whitehurst 1984). Crandall (1978) hatte (auf der Basis einer skeptischen Haltung zur Bedeutungsäquivalenz von Gutachterempfehlungen) vorgeschlagen, benachbarte Urteile einer Entscheidungsskala als übereinstimmend zu behandeln. Lindsey (1988) wandte ein, dass Crandalls 15-Felder-Matrix zur Übereinstimmung zweier Gutachter so zwar (bei Aggregierung bestimmter Zellen) eine hohe Übereinstimmungsquote, nach Abzug einer möglicherweise ‚zufälligen' Übereinstimmung aber immer noch eine extrem schwache Reliabilität zeige. Hargens/Herting (1990) kritisierten wiederum Lindseys Studie von 1979, die Übereinstimmungsmatrix der Gutachten (des *American Sociological Review*) zeige, dass die Urteile der Gutachter nicht statistisch unabhängig seien. Außerdem könne man eben nicht davon ausgehen, dass zwischen den Kategorien der Rangskala einer Publikationsempfehlung Abstandsgleichheit herrscht. Die Autoren gehen davon aus, dass die Variationen zwischen den Kategorien so stark sind, dass man die Reliabilität im Peer Review Prozess gar nicht messen kann. Lindsey (1990) insistierte auf dem kritischen Messbefund: Die Übereinstimmungsquote hängt vor allem an den 58 % Rejektionsurteilen: in der Hälfte dieser stimmen Gutachter überein. Ohne sie, und d. h. für die Aufgabe der ‚Bestenauswahl', sei die Reliabilität des Peer Review minimal.

Cicchettis gleichlautende Qualifizierung der Reliabilitätskoeffizienten als ‚schwach' stützte sich von vornherein auf die Subtraktion von Zufallseffekten von den ‚Rohdaten' der Übereinstimmungsquoten. Daraus folgt, dass etwa eine Übereinstimmung von 70 % als mangelhaft, und von bis zu 80 % als mittelmäßig eingestuft wurde. Weiter erläuterte er im Detail, dass die Wahl des statistischen Maßes davon abhängen muss, ob die gleichen oder verschiedene Gutachter Urteile abgeben und ob ihre Zahl konstant oder variabel ist.

Die messtechnische Kritik richtete sich neben der Diskussion dieser unterschiedlichen Reliabilitätsmaße vor allem gegen die Behauptung eines höheren Rejektionskonsenses. Sie beruht auf einer Matrix von Übereinstimmungsquoten verschiedener Beurteilungskategorien: Wie oft stimmt ein zweiter Gutachter mit einem als gegeben angenommenen Urteil eines ersten Gutachters überein? Der Befund scheint aber – sparsamer interpretiert (so D. Eckberg) – eher ein statistisches Artefakt der größeren Häufigkeit negativer Voten: Wenn häufiger ablehnend geurteilt wird, kommt es auch häufiger zu einer Übereinstimmung solcher Urteile. Cicchetti korrigierte zwar die allgemeinen, nicht aber die kategorienspezifischen Konsensquoten um zufällige Übereinstimmungen. Tut man dies, sind die Übereinstimmungsquoten identisch für *‚accept'* und *‚reject'*. Auch Gerald Wasserman meinte, der Rejektionskonsens könne nur die Kehrseite des Akzeptanzkonsenses sein. Ihre Differenz sei einfach eine Funktion der Akzeptanzquote einer Zeitschrift: Wenn kaum etwas akzeptiert wird, geht auch der Akzeptanzkonsens gegen Null, der Rejektionskonsens steigt. Cicchetti stimmte diesen Einwänden für den Fall der kritisierten Vierfeldertafel zu, bei mehr Kategorien (wo unterschiedliche Grade zufallskorrigierter Übereinstimmung möglich sind) ließe sich ein höherer Rejektionskonsens aber durchaus zeigen.

Ob die vorgefundene Gutachter-Reliabilität nun als ‚hoch' oder ‚niedrig' gelten soll, wurde aber nicht nur mithilfe solcher stochastischer Argumente diskutiert. Es gab auch eine Vielzahl substantieller Einwände gegen Cicchettis Diagnose einer Reliabilitätsschwäche des Peer Review: Der erste lautet, dass man von Urteilen im Peer Review nicht zuviel Reliabilität erwarten darf. Ihre Einschätzung als ‚schwach' stützt sich auf die psychometrische Perspektive der klassischen Testtheorie, die zwei Urteile als parallele Messungen einer latenten Eigenschaft betrachtet (L. Hargens). Die kognitive Psychologie hat aber in vielen Studien eine schwache Reliabilität von komplexen Entscheidungen aufgezeigt: etwa bei Einstellungen, ärztlichen Diagnosen oder Aktienkäufen. Dies ist also keine Besonderheit des Peer Review (so H. Roediger und P. Cohen). Ferner sind hier Kontextaspekte zu beachten: die redaktionelle Vorauswahl von Manuskripten (bei Top-Journals bis zu 50 %) sortiert jene indiskutablen Manuskripte, über die großer Konsens besteht, von vornherein aus, – eine Verengung des Qualitätsspektrums, die die nachfolgende Gutachterübereinstimmung erheblich reduziert (Hargens, Marsh/Ball). Man kann an dieser Stelle auch auf die Einwände verweisen, die Sandra Scarr schon gegen die Studie von Peters und Ceci (Harnad 1982) vorbrachte: Urteile über wissenschaftliche Manuskripte verlangen die Gewichtung unzähliger Kriterien. Sie sind so komplex wie Urteile über Attraktivität, über den Geschmack von Wein oder das

Aroma von Parfums. Zwar kann man gravierende Mängel recht gut spezifizieren (daher die höhere Übereinstimmung bei Ablehnungen), aber man kann wie bei Bewerbern im Einstellungsgespräch so auch bei Manuskripten ganz unterschiedliche Qualifikationsprofile beschreiben, je nachdem welchen Einsatzzweck man im Auge hat. Wählt man nur die richtigen Vergleiche (so schon Roediger 1987: 251), so erscheint die Reliabilität des Peer Review als ‚gar nicht so schlecht'.

Der zweite Typus von Einwänden läuft darauf hinaus, dass auch die Konzentration auf die Übereinstimmung der Gutachterurteile die Reliabilität des Gesamtverfahrens unterschätzt. Der Zweck des Peer Review ist nicht Übereinstimmung von Gutachtern, sondern Optimierung der Publikationsentscheidungen von Herausgebern (J. Bailar, C. Kiesler). So geht den von Cicchetti isoliert betrachteten Gutachterurteilen eine Gutachterauswahl voraus, die in zwei Hinsichten reliabilitätssenkend wirken kann. Zum einen können es Gutachter mit einem unterschiedlichen Votenstil sein, insbesondere wenn Herausgeber zu einer Mischung von erfahrenen und jungen (neu zu rekrutierenden und auch zu schulenden) Gutachtern greifen. Die Maße unterschätzen dann die Übereinstimmung, weil sie tatsächliche Meinungsverschiedenheit nicht von unterschiedlicher Strenge unterscheiden können (Daniel 1993). Zum anderen können gezielt Gutachter aus verschiedenen Schulen oder zu verschiedenen Gütedimensionen eines Manuskripts bestellt werden (Hargens). Oft werden sie sogar explizit für die Produktion von Diversität und Komplementarität des Urteilens ausgewählt (C. Kiesler, H. Kraemer). Eine hohe Übereinstimmung, so Bailar, kann dann gerade indizieren, dass schlecht ausgewählt und redundante Gutachten erstellt wurden.

Ferner dürfte die Reliabilität des Gesamtprozesses höher sein als die der Gutachterurteile, weil fast alle Zeitschriften bei *‚split votes'* weitere Meinungen einholen und i.d.R. auch eigene Herausgeberurteile hinzufügen (Hargens, Marsh/Ball) – ein altes Argument von Cronbach (1981) gegen Cole (s. a. Hargens/Herting 1990b). Dabei urteilen die Herausgeber unter Nutzung nicht nur der formalen Empfehlungen der Gutachter, sondern auch ihrer inhaltlichen Kommentare als Informationsquelle. Z.B. können die Kommentare von Gutachtern exzellent sein, von denen man aber weiß, dass sie es i.d.R. nicht schaffen, ein Manuskript abzulehnen bzw. zu akzeptieren (Bailar). Aus all diesen Gründen dürften die Herausgeberentscheidungen um vieles verlässlicher sein als das Votum des einzelnen Gutachters (Marsh/Ball).

Der dritte Typ von Einwänden bezieht sich auf die Einseitigkeit der Werteprämissen Cicchettis: Reliabilität kann – auch wenn man den Gesamtprozess im Blick hat – kein exklusives Gütekriterium von Peer Review Verfahren sein, kein ‚Selbstzweck'. Helena Kraemer warnt, dass man Reliabilität nicht auf Kosten von Validität steigern darf. Man kann ein hochreliables (präzises) Urteil ohne jede Validität haben. Es ist dann nicht mehr als ein gut reproduzierbarer Irrtum. Auch Peter Schönemann und Charles Kiesler gehen von einer Normalität von ‚Fehlurteilen' aus. Wenn Gutachter aber irren, wäre die Konvergenz ihrer Irrtümer noch verhehrender als eine bloß zufällige Streuung ihrer Voten. Es könnte sein, dass die schwache verbleibende Reliabilität in den Sozialwissenschaften vor allem auf den *confirmatory bias* (Mahoney 1977) zurückgeht, mit dem ein Feld seine Irrtümer fortschreibt. Wenn die Validität des Peer Review aber gering ist, muss man seine Reliabilität senken, nicht steigern.

Vor allem in Bezug auf diese Einschätzung von Reliabilität als Gütekriterium stehen sich in der Peer Review Forschung recht unversöhnliche Auffassungen gegenüber. Auf der einen Seite entwickelt Lindsey (1988) aus dem Befund eines schwachen Akzeptanzkonsenses ein vernichtendes Urteil über den Peer Review. Er geht davon aus, dass sich dessen Funktion historisch verschoben hat: Ging es in den frühen Zeitschriften darum, Unsinn auszuscheiden, so geht es heute in vielen Zeitschriften – bei zahllosen Manuskripten und höheren Rejektionsquoten – darum, die 10–15 % besten Manuskripte auszuwählen. Wenn die Übereinstimmung von Gutachtern nun aber vorwiegend bei den Rejektionsurteilen besteht, sinkt sie dramatisch, wenn man die abgelehnten Manuskripte subtrahiert. Lindsey folgert: Die meisten Publikationen in den Sozialwissenschaften (wo hohe Rejektionsraten herrschen) stammen aus Entscheidungsverläufen mit minimaler Reliabilität. Die Differenz zur zufälligen Übereinstimmung wird so klein, dass Peer Review Verfahren nur „etwas besser als ein Würfel" abschneiden. Auch Bornstein (1991) resumiert: Wenn man den Peer Review wie bisher üblich als quasi-empirisches Verfahren betrachtet, so muss man feststellen, dass er in allen möglichen Aspekten wissenschaftlicher Güte versagt: in verschiedensten Reliabilitäts- und Validitätsmaßen. Der Peer Review wäre selbst nach den mildesten wissenschaftlichen Standards für Assessment-Instrumente inakzeptabel: Würde man ein Manuskript mit Daten veröffentlichen wollen, deren Reliabilität so schwach wie die des Manuskript-Auswahlverfahrens wäre, so wären seine Publikationschancen gleich Null. Singer (1989) spricht schlicht von einer ‚Kriterienkrise' der Wissenschaft.

Auf der anderen Seite finden sich Bewertungen, die im Gefolge der genannten Argumente auf Entdramatisierung drängen. Schon Cole/Cole/Simon (1981) sehen die schwache Reliabilität von Peer Review Verfahren als eine unproblematische Meinungsverschiedenheit. Sie folge einfach aus dem schwachen kognitiven Konsens an den Forschungsfronten aller Disziplinen (Cole 1983).Vor allem mit dieser Tatsache müsse man umgehen lernen. Über diese Position hinausgehend kritisiert Harnad (1985) eine auch noch bei Cole vorfindliche Wertschätzung von Konsens, die dazu tendiert, Meinungsverschiedenheit als *‚random'* und wissenschaftliche Fehlleistung abzuwerten. Konsens, so Harnad, ist etwas für die Wissenschaftsgeschichte, an den Forschungsfronten herrschen Versuch und Irrtum, Vermutung, Zufall und Wettbewerb, kurz: „*creative disagreement*".

2.3 Inhaltsanalysen von Gütekriterien

Im Gegensatz zur technischen Intervallmessung von Eckdaten des Peer Review bemühte sich eine kleine Gruppe von Studien um eine stärkere Annäherung an den Prozess der Urteilsbildung. Mit Hilfe von Dokumentenanalysen der Fachgutachten versuchte man, den Kriterien und Entscheidungsgründen der Reviewer auf die Spur zu kommen (etwa Smigel/Ross 1970, McCartney 1979, Fiske/Fogg 1990, Bakanic/McPhail/Simon 1989).

Smigel/Ross (1970) bildeten für den Fall der Zeitschrift *Social Problems* ein Ranking von Gründen der Akzeptanz oder Ablehnung und zeigten, dass unterschiedliche Publikationsempfehlungen aus verschiedenen Gewichtungen von Kriterien resultieren. Fiske/Fogg (1990) bestätigten den Befund, dass zwei Gutachter (zum Leidwesen der Autoren) völlig verschiedene Punkte ansprechen und aus diesem Grund divergierende Empfehlungen geben. Bakanic/McPhail/Simon (1989) analysierten 800 Gutachten für den *American Sociological Review* anhand von verschiedenen Beurteilungskriterien wie Einschlägigkeit, Literaturrezeption, Originalität, methodische Korrektheit, Relevanz der Ergebnisse, Klarheit der Darstellung usw. Auch sie stellten fest, dass Publikationsempfehlung und Manuskript-Beurteilung unabhängig voneinander variieren: Übereinstimmende Verfahrensvorschläge von Gutachtern können auf völlig verschiedenen Gründen basieren und übereinstimmende inhaltliche Einschätzungen eines Manuskripts zu ganz verschiedenen Publikationsempfehlungen führen, je nachdem wie die Gutachter ihre Einwände gewichten. 11 % der Manuskripte bekamen explizit widersprüchliche Kommentare, aber 40 % unterschiedliche Publikationsempfehlungen. Andererseits gab es eine hohe Übereinstimmung in der Valenz der einzelnen Kategorien. Der häufigste Fall war aber, dass die Gutachter sich mit ihren Urteilen einfach auf verschiedene Aspekte des Manuskripts bezogen. Außerdem ließen sich Unterschiede in der Übereinstimmung je nach Kategorie feststellen: Dissens war bei Fragen der Interpretation und Theorie höher als bei methodischen Fragen.

Auch andere Studien (etwa: Hartmann/Neidhardt 1990) stellten fest, dass sich der hohe Kriterienkonsens, den man mittels einfacher Befragungen erheben kann, in der praktischen Anwendung von Kriterien schnell auflöst. Redaktionell vorgegebene Kriterienkataloge scheinen vor allem daran zu scheitern, dass Aspekte wie ‚Innovativität', ‚methodische Korrektheit' und ‚Lesbarkeit' in Zielkonflikte geraten können (Beck/Hartmann 1983: 266). Cicchetti/Eron (1979) stellten gar fest, dass die Reliabilität der Urteile mit der Verwendung von Kriterienkatalogen noch abnimmt. Versucht man die Urteile mittels Formalisierung zu steuern, scheint man ihre Divergenz eher zu erhöhen.

Ein weiteres Thema inhaltsanalytischer Studien schneidet eine Untersuchung von Neidhardt (1986, 1988) über Forschungsanträge an. Der auffälligste Befund ist hier eine ‚argumentative Inkonsistenz', mithilfe derer die Gutachten erst ihre Urteilstendenz bekommen: Obwohl sich die Gutachter gerade in den zentralen Kriterien wissenschaftlicher Güte (Theorie und Methode) am kritischsten äußern, lassen sie dies doch nur selten in eine Ablehnungsempfehlung münden. Diese ‚Milde' lässt sich nicht einfach statistisch mit der Überlagerung durch andere Kriterien (Machbarkeit, Kosten etc.) erklären (1988: 114). Neidhardt gibt stattdessen eine soziologisch-theoretische Erklärung, die auf *Kollegialität* rekurriert: ein „berufsspezifisches Subkulturmuster, das Solidarverpflichtungen der Mitglieder zueinander normiert" (1986: 6). Im Fall von Forschungsanträgen stecken die Gutachter im sozialen Dilemma einer Doppelbindung an die Aufraggeber einerseits, die Fachkollegen andererseits, denen gegenüber sie zugleich *Peer* und Statusrichter (für einen Dritten) sein sollen. Die daraus resultierende ‚Beißhemmung' schlägt sich in einer Reihe von rhetorischen Maßnahmen nieder, mit denen sich die Gutachter bei der Formulierung ihrer Kritik um Entschärfung und Schadensbegrenzung bemühen: Sie wählen schonende, konjunktivische Formulierungen, relativieren ihr Urteil zur Meinungsverschiedenheit und bereiten schon in der Kritik Rückzugspositionen im Hinblick auf die Förderungsempfehlung vor.

Neidhardts Studie macht auf zwei Desiderate aufmerksam: zum einen auf den Theoriebedarf im Hinblick auf Sozialbeziehungen zwischen den Teilnehmern des Peer Review.[9] Zum anderen zeigt die Studie aber auch die methodischen Grenzen einer *Inhalts*analyse von Gutachten. Neidhardt selbst verweist auf das Problem, dass eine Nicht-Nennung von Gründen keine Nicht-Berücksichtigung impliziere (1988: 94). Darüber hinaus steht der Aspekt der Rhetorik aber nicht nur für Formfragen und stilistische Bemäntelung (aus linguistischer Perspektive etwa Kretzenbacher/Thurmair 1992, Johnson 1992, Johnson/Roen 1992, He 1993), sondern für die *Praxisdimension* der schriftlichen Darstellung von Güteurteilen. Spencer/Hartnett/Mahoney (1986) versuchten noch in einer objektivistischen Lesart von Gutachten „substantielle Argumente" von „bloßer Rhetorik" (von Emotionalem und Persuasivem) zu unterscheiden. Gegen diese Unterscheidung sind im Kontext der Peer Review Forschung aber zwei Einwände aufgebracht worden, die dafür sprechen, Fragen der Rhetorik weit ernster zu nehmen als in Inhaltsanalysen möglich.

Zum ersten entnahm Cicchetti (1991) Studien über die Inkonsistenzen zwischen Kriterienvalenz und Publikationsempfehlung einen Hinweis auf die Psychologie der Urteilsbildung: Es sei gut möglich, dass Gutachter sich zuerst für eine Publikationsempfehlung entscheiden und dann ihr Gutachten verfassen und einzelne Kriterien veranschlagen. Man kann diese Hypothese auch ethnomethodologisch formulieren: Die Bewerkstelligung des ‚Urteilens' besteht eben auch in der rhetorischen Herstellung von Konsistenz durch die Legitimation einer einmal getroffenen (Vor)entscheidung. Dabei sind die an der Textoberfläche erscheinenden Gründe post hoc Rationalisierungen, die mit den im Entscheidungsprozess wirksamen Gründen und Motiven u. U. nicht viel zu tun haben.

Ein zweiter Hinweis tauchte in der Diskussion der Studie von Peters und Ceci auf. Diese waren Donald Beavers Vorhalt, ihre gefakten Manuskripte seien wohl implizit als ‚veraltet' abgelehnt worden, mit dem Argument begegnet, dass die Gutachter in keinem Fall fehlende Innovativität, aber in fast allen Fällen (eher ‚zeitlose') „methodische Mängel" konstatierten. Beaver hielt dagegen, dass der Eindruck fehlender Originalität sehr wohl eine zentrale Rolle spielen kann, ohne dass er auch als Entscheidungsgrund mitgeteilt wird: Der Hinweis auf „methodische Fehler" hat große argumentationsökonomische Vorteile gegenüber der disputablen Feststellung von Redundanz. Auch William Honig und David Palermo verwiesen auf eine Differenzierung zwischen den Kriterien der Beurteilungspraxis und den (inhaltsanalytisch feststellbaren) Beurteilungsgründen in den Gutachten. Das Argument, etwas sei ein ‚alter Hut', wird nur sehr selten in den Gutachten vorgebracht, da es oft ‚Argumentationsfolgekosten' – detaillierte Entgegnungen und Nachfragen – nach sich zieht. Einfacher ist es, sich auf andere Mängel des Manuskripts zu beziehen, vorzugsweise eben methodische. Auf diese Präferenz verwies schon Mahoneys experimentelle Studie von 1977, die mit einer Fingierung von theoretischen Positionen arbeitete: Bestätigten die Daten vorherrschende theoretische Ansichten, fanden nur 25 % der Gutachter „methodische Mängel", widersprachen sie ihnen, waren es 71 %. Wenn man Bakanic/McPhail/Simons (1989) Befund hinzunimmt, dass auch der Gutachter-*Konsens* bei Methodenkriterien höher ist als bei anderen, so kann man vermuten, dass „methodische Schwächen" eine *zustimmungs*suchende *Ablehnungs*rhetorik sind.[10]

2.4 Zitationsanalysen zur Vorhersagevalidität

Während Inhaltsanalysen die Tiefenschärfe der Peer Review Forschung verbesserten, bemühten sich andere Untersuchungsstrategien um eine Verbreiterung der Perspektive, indem sie einzelne Zeitschriften im Kontext eines Marktes betrachteten und die ‚Rezeptionszukunft' eines Manuskripts zur Validierung von Publikationsentscheidungen heranzogen. Sie nutzten dabei mit Zitationsanalysen ein anderes Evaluationsverfahren (dessen Daten auf den Ergebnissen des Peer Review beruht) für eine Apologetik des Peer Review. Wo Inhaltsanalysen auf den Gutachter als Textproduzenten aufmerksam machen, haben solche Studien den Autor als Publikations-

[9] Hier ist in der Manuskriptbegutachtung neben der Kollegialität sicher auch an Konkurrenz und Patronage zu denken.

[10] Ein Reviewer dieses Aufsatzes wendet hier ein: „die Zentrierung auf methodische Mängel erfüllt nicht nur eine argumentationsökonomische Funktion für Gutachter, sondern erbringt auch einen sachlichen Gewinn". Als Fachwissenschaftler kann man hier nur beipflichten, dass die *Nennung* methodischer Schwächen Manuskripte optimiert. Als wissenschaftssoziologischer Beobachter kann man aber auch sehen, dass die *Zentrierung* auf methodische Schwächen eben doch primär rhetorische Gründe hat. Dann kann man als Fachwissenschaftler wiederum etwas besser verstehen, warum man in Zeitschriften soviele „methodisch saubere, aber sachlich langweilige und theoretisch belanglose Artikel" (Burkart 2002: 49) zu lesen bekommt.

strategen im Blick.[11] In der pointierten Formulierung von Tyrer (in Cicchetti 1991): „A determined author can get any rubbish published".

Die gemeinten Studien stützen sich auf den Umstand, dass ein Großteil abgelehnter Manuskripte später in anderen Zeitschriften zur Publikation kommt (Wilson 1978, Garvey 1979). Herausgeber besonders reputierter Zeitschriften berichten, dass sie zwischen 68 % und 100 % abgelehnter Manuskripte (oft unverändert oder nur marginal überarbeitet) später in anderen, durchaus angesehenen Zeitschriften wiederfanden. Rourke (in Cicchetti 1991) schildert den Fall eines Manuskripts, das ihm binnen 14 Monaten von vier verschiedenen Zeitschriften unverändert zur Begutachtung zugesandt wurde.

Eine Überprüfungschance für die Validität von Publikationsentscheidungen sehen manche Autoren nun darin, dass sich Zeitschriften über ihre Zitationsquoten in eine Rangordnung bringen lassen. Vor allem Herausgeber medizinischer Zeitschriften argumentierten so für die Qualität ihrer Verfahren: Wilson (1978) demonstrierte, dass die zur Publikation akzeptierten Manuskripte eines Jahrgangs doppelt so häufig zitiert wurden wie abgelehnte, die anschließend in anderen Zeitschriften erschienen. Lock (1985) zeigte, dass von den 79 % abgelehnten Manuskripten des *British Medical Journal* nur 16 % in Zeitschriften mit gleichem oder höherem *Impact Factor* (der Gesamtzitationsquote der Zeitschrift) erschienen. Stossel (1985) berichtet für das *Journal of Clinical Investigation* (JOCI), dass die Zitationsrate der biomedizinischen Zeitschriften, in denen sich abgelehnte Manuskripte des JOCI wiederfinden, etwa bei 50 % der des JOCI lag. Für die Zeitschrift Angewandte Chemie verfolgte Daniel (1993) die 71 % der abgelehnten Manuskripte, die später in anderen Zeitschriften erschienen. Auch hier erschienen alle in Zeitschriften mit niedrigerem *Impact Factor*. Da dieses Maß insofern grobschlächtig ist, als der *Impact Factor* erheblich durch herausragende Einzelpublikationen bestimmt werden kann, verfolgte Daniel auch noch jedes einzelne Manuskript. Das Ergebnis: abgelehnte Manuskripte erhalten im Schnitt nur eine halb so hohe Zitationsrate.

Gegen diese Untersuchungsdesigns lassen sich ebenfalls eine Reihe von Einwänden vorbringen. Daniel selbst erwägt den Einwand, dass die höhere Zitationsquote schlicht durch das Erscheinen in einem sog. ‚*high impact journal*' hervorgerufen werden könnte (1993: 74). Die Publikationsentscheidung bestimmter Zeitschriften hebt eben schon die Zitationswahrscheinlichkeit einzelner Beiträge. Weitere Einwände könnten gegen die erhebliche Selektivität des Datensatzes des *Social Science Citation Index* (Fröhlich 1999), gegen die theoretisch ungeklärte Bedeutung von Zitaten in der Wissenschaftskommunikation und gegen das unklare Verhältnis von Lektüre und Zitat erhoben werden. Beide dürften sich quantitativ weit überschreiten: Es wird viel mehr gelesen als zitiert und es wird Vieles als ‚zitierpflichtig' behandelt, ohne dass es gelesen wurde.

Es gibt aber noch einen einfacheren Grund, die hartnäckige Suche nach einem solchen ‚unabhängigen Maß' wissenschaftlicher Güte für sinnlos zu halten. Die absoluten Zitationsziffern, die sich hinter den Quoten und Rankings verbergen, sind nämlich so gering (bei der Angewandten Chemie waren es im Schnitt in 5 Jahren 12 Zitate, bei den anderen Zeitschriften nur 6, also pro Jahr eine Differenz von einem Zitat), dass man diese Zitate eigentlich nur als ‚ein weiteres Urteil' betrachten kann, das Leser nach den Herausgebern und Gutachtern über ein Manuskript abgegeben haben. Dabei kann kaum beeindrucken, dass ein guter Teil der Manuskripte im Laufe der Jahre mehr zitierende als gutachtende Leser findet. Entscheidend ist, dass dieses Maß gar nicht darüber informieren kann, ob das Urteil dieser zitierenden Leser nun zustimmend, ablehnend oder gänzlich indifferent ist. Die Aggregation von Zitationsquoten suggeriert einen ‚Schiedsrichter Markt', der eine editorische Vorselektion bestätigt oder zurückweist. Für den weitaus größten Teil aller Zeitschriftenaufsätze gibt es anstelle eines solchen Marktes aber nur ein punktuelles Anschließen von Zitaten an Aufsätze (und andere Zitate), das nur an die Urteile von Gutachtern und Herausgebern anschließt. Ferner ist davon auszugehen, dass bei weniger renommierten Zeitschriften die Mehrzahl der Manuskripte *mehr* Leser *im* Peer Review als auf dem ‚Markt' finden werden.[12] Kurz:

[11] Auf diesen zielen auch zahlreiche „How to Publish"-Ratgeber, etwa Thyer 1994, Day 1996 und 1998, Hall 1998.

[12] Lesererhebungen gehören ebenfalls zu den Desideraten der Wissenschaftsforschung. Das letzte Editorial dieser Zeitschrift berichtete über eine kleine Erhebung der Leserschaft von 8 Artikeln, die kurz vor der Umfrage erschienen. „Ganz oder weitgehend" gelesen wurde zwischen 9 und 27 % (im Schnitt 18 %), „teilweise" zu weiteren 8 %. Diese Zahlen sprechen einerseits für eine recht aktive Leserschaft, sind andererseits aber auch mit Vorsicht zu genießen: 1. dürfte der Fragebogenrücklauf von 37 % jene Abonnenten angesprochen haben, die sich der Zeitschrift auch durch Lektüre verbunden fühlen. Dazu zählt dann aber auch eine Empfänglichkeit für die soziale Erwünscht-

Eine Quantifizierung macht bei diesen Fallzahlen wenig Sinn. Sie scheint eher motiviert durch das schiere Vorhandensein solcher Datensätze: eine „Abfallforschung“ (Fröhlich 1999), die das leicht Messbare mit dem theoretisch Relevanten verwechselt.

3. Konzeptuelle Schwächen der Forschung

Mein knapper Review der Peer Review Forschung hat neben der Darstellung zentraler Studien auch bereits eine Reihe von kritischen Einwänden aufgelistet. Die wesentlichen Schwächen der Peer Review Forschung liegen aber nicht in den Details der Untersuchungsdesigns und der Messtechnik, sie liegen im Grundsätzlichen: in den Erwartungshaltungen der Forschung sowie in einem Professionalitäts- und Theoriedefizit. Die Kritik muss daher tiefer ansetzen und viel stärker als bisher auf *Prämissen* der Forschung gerichtet werden.

Die Theorieschwäche der Peer Review Forschung ist schon an ihrer normativen Überdetermination erkennbar: Die Forschung ist z. T. hoffnungslos evaluativ, befangen zwischen Kritik und Apologetik. Dabei sind auch Probleme der Rollendifferenzierung erkennbar: Mitunter hat man den Eindruck, die vernichtendste ‚Sozialkritik‘ des Peer Review stammt von enttäuschten Autoren, die rationalistische Apologetik von Herausgebern einer Zeitschrift.[13] Besonders in den experimentellen Studien herrscht ein Tenor des *‚blaming the reviewers‘* vor, die Perspektive von Autoren also: über fallweise unfaire, in jedem Fall aber ‚unwissenschaftliche‘ Entscheidungsprozesse.

Der Stand der Forschung zeigt insofern, wie schwierig es ist, gerade bei der Evaluation von Evaluationsverfahren von einer szientistischen und normativen Selbstbeobachtung von Wissenschaft zu einer professionellen Wissenschaftsforschung überzugehen, die soziale Prozesse als immanenten Bestandteil der Leistungsfähigkeit des Wissenschaftssystems sieht. Die Peer Review Forschung ist zu großen Teilen Amateur-Wissenschaftsforschung, die von Wissenschaftlern betrieben wird, die nicht für die Beobachtung ihrer eigenen Praxis ausgebildet wurden.[14] So wiederholt man oft gebetsmühlenhaft die ‚Erkenntnis‘, dass Wissenschaftler im Peer Review *nicht* wissenschaftlich arbeiten, und verweigert mit eben dieser Wiederholung die Reflexion darauf, dass dies nicht primär an deren individueller Fehlbarkeit, sondern an gänzlich deplazierten Vorstellungen der Untersuchenden von wissenschaftlicher Praxis liegen könnte.

Der Peer Review Forschung fehlt es also generell an einer professionspolitischen ‚Abkühlung‘ im Sinne von Wissenschafts*forschung*. Anstelle des normativen Pro und Contra sind zunächst eine Vielzahl empirischer und analytischer Fragen zu stellen, mit denen sich eine Reihe fragwürdiger Prämissen überprüfen und ersetzen lassen. Eine solche Hinwendung zu Prämissen der Forschung wurde bereits 1982 von Mahoney (in der Diskussion mit Peters und Ceci) gefordert. Er stellte fest, dass das Konzept der Reliabilität genauso wie das der Objektivität davon ausgeht, man könne die Wirklichkeit einer Sache unabhängig von menschlichen Wissensprozessen bestimmen. Nimmt man davon Abstand, erscheint die Peer Review Forschung weniger als engagierte Anklage gegen Fehlleistungen denn als unfreiwillige Aufklärung über latente Annahmen über wissenschaftliches Wissen, z. B. das Verhältnis zu akademischer Autorität. Die wesentliche Aufgabe, so Mahoney, besteht nicht in der Verbesserung des Systems, sondern darin, uns seine Funktionen besser klarzumachen: Wie wird Wissen akkreditiert und verbreitet?

3.1 Prozesse statt Personen

Gehen wir im Versuch einer auf Prämissen und Common Sense Annahmen zielenden Kritik zu-

heit lektürefleissiger Antworten. 2. Aussagekräftiger als eine Aggregation zu durchschnittlichen Prozentangaben sind a) die ungleichen Verteilungen auf einzelne Artikel: es gibt viel und gar nicht gelesene b) die absoluten Zahlen: und diese können in einzelnen Fällen eben die der Peer Review Beteiligten leicht unterschreiten. 3. nicht nur die Auskünfte „teilweise“ und „weitgehend“ gelesen, sondern auch die Selbstauskunft „gelesen“ kann auf eine weit oberflächlichere Form der Lektüre verweisen als die der Peer Review Teilnehmer, die ihre Leseeindrücke zu schriftlichen Stellungnahmen machen müssen, die auch von anderen beobachtet werden (s. u.).

[13] Das ist psychologisch durchaus verständlich: In einer gänzlich opportunistischen Wissenssoziologie leuchten uns als Autoren ‚kognitive‘ Gründe einer Herausgeberentscheidung immer dann ein, wenn unsere Manuskripte akzeptiert werden; werden sie abgelehnt, erkennen wir schlagartig die ‚soziale‘ Dimension des Peer Review.

[14] Dies gilt besonders für die im JAMA dokumentierte medizinische Forschung. Oft handelt es sich nur um objektivistische ‚Falsifikationen‘ von ‚peer reviewed‘ Artikeln, die Meinungsverschiedenheiten über angeblich fraglose Standards einer Disziplin austragen – eine Fortsetzung von Fachdebatten im (fadenscheinigen) Gewand der ‚Wissenschaftsforschung‘.

nächst auf die experimentellen Studien zum Gutachterbias ein. Ihre Fragestellungen sind schon dadurch beschränkt, dass sie sich unter Vernachlässigung der Prozesslogik wissenschaftlicher Kritik (wie sie schon seit den Arbeiten von Popper im Zentrum der Wisenschaftstheorie stehen) voll und ganz auf Personen (auf Merkmale von Autoren und Gutachtern) richten. Die implizite normative Folie dieses Fokus ist eine Tugendlehre des ‚weisen, vorurteilsfreien' Forschers – eine im Grunde vordemokratische Erwartung. P. Abelson (1980: 62) widmete ihr bereits 1980 einen sarkastischen Kommentar: „But why peer review? Why not an objective, allknowing, all-wise genius to serve as editor?" Seine Antwort: „Such mortals do not exist."

Man kann an die Stelle dieser Erwartung zunächst eine Reihe von einfachen Prozess-Anforderungen setzen, an denen sich Peer Review Verfahren zu messen haben, darunter die Anonymisierung, eine pluralistische und temporäre Besetzung von Entscheidungsrollen, die Sicherung einer gewissen Anzahl von Stimmen bzw. einer Mehrinstanzlichkeit des Entscheidungsprozesses (insbesondere in multiparadigmatischen Disziplinen) usw.

Innerhalb der Diskussion der Studie von Peters und Ceci wurde der Fokus auf individuelle Personen immerhin insofern korrigiert, als mit dem Argument der adressatenspezifischen Überarbeitung der Manuskripte für seine *Erst*begutachter weitere Personen in den Blick genommen wurden. Ein Manuskript kann schon deshalb nicht gut mit individuellen Merkmalen des Autors identifiziert werden, weil es durch seinen adressatenspezifischen Zuschnitt zu wesentlichen Teilen ein gemeinsames Produkt von Autor und Reviewer ist. Bös (1998) assoziiert in diesem Sinne Peer Review Zeitschriften mit dem Aufschwung der Koautorschaft: „Wissenschaftshistoriker werden später einmal Schwierigkeiten haben, zu beurteilen, welche Ideen von den Autoren kommen und welche von… Referees" (ebd.: 71). In der Tat gibt es ein gewisses Missverhältnis zwischen der geläufigen Problematisierung eines ‚geistigen Diebstahls' durch Gutachter, die einen Autor durch eine Kombination von ‚Hinhalten und Abkupfern' um den gerechten Lohn bringen, und der Thematisierung der Frage, wie man in einem kollektiven Wissensprozess die Beiträge jener ‚aktiven Leser' honoriert, die bei einer Publikation gar nicht als Autoren in Erscheinung treten. Ein Zeitschriftenaufsatz ist ja nicht nur eine Information, er ist auch eine Urkunde, die einen kommunikativen Akt dauerhaft als ‚geistiges Eigentum' zuschreibt. Wichtiger als die besitzrechtliche Frage der gerechten Reputationsverteilung ist hier freilich die analytische Frage, ob man Publikationen nicht als hochartifizielle Individualisierungen kollektiver Forschungsprozesse betrachten muss. Publikationen gehören zu den kommunikativen Akten, mit denen die Wissenschaft ihre ‚Individuen' erst *herstellt*.

Gravierender noch ist ein weiteres Problem der experimentellen Studien. Die Diskussion der Studie von Peters und Ceci zeigte zwei problematische Konstanzannahmen des experimentellen Designs: die Annahme einer ‚Alterslosigkeit' von Manuskripten und die Abstraktion von der Perzeption einer *Wieder*einreichung als *Erst*einreichung. Hinter diesen Konstanzannahmen steht eine implizite Ausdehnung der Eigenschaftspsychologie von Personen auf Manuskripte. Bei der konzeptuellen Erfassung von Manuskripten reicht es aber nicht (wie in der Diskussion um Cicchetti), nur die *Komplexität* des Beurteilungsgegenstands hervorzuheben. Denn kommunikative Akte, um die es sich bei Manuskripten (und ihren Einreichungen) handelt, lassen sich überhaupt nicht ohne weiteres als ‚Gegenstände' fassen. Smigel/Ross (1970: 21) meinten in ihrer Inhaltsanalyse von Fachgutachten, es sei ihnen manchmal schwer gefallen zu glauben, dass die Gutachter dasselbe Manuskript gelesen hatten. Eben dies kann man aus einer rezeptionstheoretischen Perspektive (Iser 1972) auch für höchst unwahrscheinlich halten: Wie kann man annehmen, dass verschiedene Leser bei einer mehrstündigen ‚Äußerung' *dasselbe* zur Kenntnis nehmen?

Beck/Hartmann (1983) sprechen in diesem Sinne davon, dass die Eigenschaften eines Manuskripts *relational* sind, d. h. sie konstituieren sich in der Beziehung zwischen den Absichten des Autors und den Erwartungen des Lesers. Seine Eigenschaften werden nicht bloß in einer statisch konzipierten Sozialbeziehung von Autor und Gutachter (Unterschiede von Rang, Alter, Geschlecht etc.) ‚richtig oder falsch' eingeschätzt, sie *entfalten* sich vielmehr erst in einer dynamischen Kommunikationsbeziehung, in der ein Autor Ansprüche erhebt und enttäuscht, Sympathien und Antipathien weckt, und ein Leser Erwartungen und ganz unterschiedliche Nutzungsinteressen im Kontext seiner eigenen Arbeiten mitbringt. Dieses Argument wurde in der Diskussion um Peters/Ceci an der Stelle gestreift, wo man vermutete, das Forschungsdesign sei unzulässig reaktiv, weil Gutachter auf starke Claims von ‚No-Names' anders reagieren als von Reputierten. Diese ‚Reaktivität' ist aber nicht einfach eine methodische Schwäche der Peer Review Forschung, sie ist vielmehr eine grundlegende Eigenschaft kommunikativer Akte, der eine Reifikation von Manuskripten als ‚Messgegenstand' nicht gewachsen ist.

Über dieses soziologische Argument hinaus ist die Konzeption von Manuskripten als mit sich identischen Gegenständen aber auch historisch unangemessen. Cicchettis Argument, dass auch ein einmütig akzeptiertes Manuskript nach seiner Publikation eine Kritik evozieren kann, deren frühzeitige Mobilisierung die Publikation hätte verhindern können,[15] verweist auf ein grundlegendes Problem objektivistischer Forschungsdesigns mit der Temporalität von Wissensprozessen: Die wissenschaftliche Güte eines Manuskripts ist wesentlich und notwendig eine historische Variable. Den ‚alterslosen' Manuskripten der experimentellen Peer Review Forschung fehlt das, was sie erst zu bestätigungs- oder kritikfähigen Kommunikationsangeboten macht: ihr historischer Index.

3.2 Meinungsbildung statt reliable Messung

Auch die Diskussion von Studien zur Reliabilität des Peer Review hat zentrale Defizite der Forschung erst gestreift. Dazu gehören zwei grundsätzliche methodisch-semantische Probleme der Reliabilitätsmessung: 1. das Problem der Bedeutungsäquivalenz: Die (mathematisch kodierbaren) Empfehlungskategorien werden von den Gutachtern idiosynkratisch und mit verschiedenen Bedeutungen verwendet und diese Bedeutung kann auf verschiedenen Dimensionen angesiedelt sein (die Zeichen markieren z. B. auch die Nachdrücklichkeit bzw. Unsicherheit des Urteils, sind also Teil einer rhetorischen Praxis). 2. das Problem der Abstandsgleichheit: In der Forschungspraxis werden Ordinalskalen häufig als Intervallskalen behandelt; dies ist aber angesichts der ganz verschiedenen praktischen Konsequenzen, die aus einem Urteil folgen, eine zwar messtechnisch bequeme, aber sozialtheoretisch fiktive Annahme.

Ebenso richtig ist ferner das Argument, dass die Erwartungen an die Übereinstimmung bei Urteilen über wissenschaftliche Arbeiten nicht zu hoch geschraubt werden dürfen. Reliabilitätsmessungen werden üblicherweise an Ratings vorgenommen, die sich auf einfache Beurteilungsaufgaben (etwa Verhaltensklassifikationen) beziehen, die mithilfe von klaren Standards (etwa das ‚richtige Wissen', auf das sich Lehrer bei der Zensurengebung stützen) und von Operationalisierungen und Schulungen, die Rater aufeinander abstimmen, gelöst werden. Die Beurteilung wissenschaftlicher Manuskripte kann keine dieser Voraussetzungen erfüllen.

All diese Argumente verharren aber noch in einer unhaltbaren Konzeption des Peer Review als wissenschaftliches *Messverfahren*, das im Prinzip nach den Gütekriterien der quantitativen Methodologie zu beurteilen ist. Peer Review Forschung scheint häufig von einem enttäuschten Szientismus geprägt, der sich durch die starke Kopplung mit Gerechtigkeitsfragen laufend selbst dementiert. Wenn man den Dissens von Gutachtern explizit oder implizit als wissenschaftliche ‚Minderleistung' darstellt, unterhält man ein feindseliges Verhältnis zu Meinungsverschiedenheiten in ergebnisoffenen Wissensprozessen. Die Übereinstimmungserwartungen in solchen Wissensprozessen sind gewissermaßen ‚schülerhaft', weil sie verbindliche Standards auch noch dort erwarten, wo diese gerade verunsichert werden, umstritten sind und neu emergieren.

Die oben dargestellte Kritik von Reliabilitätsmessungen ist ebenso trifftig mit ihrem Hinweis auf Verfahrenskontexte: Die Querschnittsbetrachtung der Urteile von Fachgutachtern macht einen arbiträren Schnitt in einem gestuften Prozess der Meinungsbildung, in dem Herausgeber einer Zeitschrift neben ihren eigenen Urteilen als Leser ein zweites, ‚supervidierendes' Urteil zu vollziehen haben. Entscheidend ist dabei aber nicht, dass so die bloße Zahl von Urteilen (und damit auch die Reliabilität) erhöht wird, sondern dass Herausgeber kaum jemals einen rein mathematischen Gebrauch von den Gutachter-Empfehlungen machen, etwa im Sinne einer Durchschnittsbildung, sondern sie *qualitativ* evaluieren: Sie müssen ihre Argumente gewichten, ihre höchst unterschiedliche Qualität einschätzen, ihre Parteilichkeit und strengen oder milden Urteilsstile (Siegelman 1991) berücksichtigen, sowie die Erwartungshaltungen des gutachtenden Lesers mit den Absichten des Autors vergleichen. Wenn man diesen qualitativen Gebrauch von Gutachten vollständig durch eine mathematische Evaluation von Gutachten ersetzt, begeht man gegenüber der sozialen Praxis des Peer Review einen *methodischen Kategorienfehler.*[16]

[15] Roediger (1987) macht ein ähnliches Argument im Kontext der Kritik überspannter Reliabilitätserwartungen: Wenn man beliebige Leser zu ihrer Meinung zu hochzitierten Schlüsseltexten einer Disziplin fragen würde (er nennt die Experimente von Milgram), würde man ähnliche Meinungsunterschiede feststellen wie bei ihrer Beurteilung im Peer Review. Wenn Wissenschaftler aber noch nicht einmal über die Güte der einflussreichsten Arbeiten ihres Forschungsfeldes Einigkeit erzielen, ist es überhaupt nicht überraschend, dass dieser Konsens bei Manuskripten fehlt.

[16] Die dargestellten zentralen Debatten der Peer Review Forschung zeigen auch ein eigentümliches Missverhältnis

Auch die Hinweise der Diskutanten auf den Kontext des Gesamtverfahrens müssen von der Wertprämisse der Steigerung von Reliabilität gelöst werden. Eine Gutachterauswahl, die explizit Meinungsverschiedenheit sucht und sich auf Varietät von Urteilen stützt, erklärt nicht bloß schwache Reliabilitätskoeffizienten, sie verweist vielmehr darauf, dass die wissenschaftliche Urteilsbildung über Manuskripte ganz anders konzipiert werden muss als ein Messvorgang. Ein sozialer Prozess, der Meinungsverschiedenheit und Perspektivenvariation *braucht*, ist von vornherein fehlkonzipiert, wenn man ihn nach dem Muster der Eichung von Instrumenten begreift. Eine solche Laborperspektive ist einfach deplaziert, wenn es um andere wissenschaftliche Praktiken geht: etwa um Lehrtätigkeit, Schreibtätigkeit oder eben editorische Urteilsbildung.

Auf der anderen Seite gibt es natürlich eine mögliche theoretische Relevanz von Reliabilitätsmessungen: die Assoziation von Reliabilität mit disziplinärem Konsens. Eben dies war schon die Annahme in den Studien von Zuckerman/Merton und Cole et al. Aus den Ergebnissen der Inhaltsanalysen lässt sich nun aber entnehmen, dass die Übereinstimmung zweier oder mehrerer Gutachter nicht mit ‚Konsens' zu verwechseln ist (sondern oft nur aus der *Konvergenz* ganz unterschiedlich begründeter Urteile folgt), und dass ihre Nicht-Übereinstimmung nicht ‚Dissens' bedeuten muss (sondern nur die differenzielle Beachtung und Gewichtung unterschiedlicher Güteaspekte). Will man die Frage nach disziplinärem Konsens empirisch differenziert stellen, wird man untersuchen müssen, wer oder was in den unterschiedlichen Phasen des Peer Review eigentlich Konsens *braucht*? Autoren, die stark divergierende Gutachten bekommen, brauchen u. U. editorische Instruktionen zur Gewichtung der Urteile. Herausgebergremien, die mit Panelentscheidungen arbeiten, brauchen einen gewissen Konsensgrad, weil Interaktionen (und Personen in *face to face* Situationen) nur eine begrenzte Verarbeitungskapazität für Dissens haben (Bohn 1999: 91). Aber schriftlich urteilende Gutachter, die weder mit den Autoren noch untereinander in Kontakt treten, haben diesen Konsensbedarf nicht.[17]

3.3 Sprech- und Schreibpraxis statt Kognition

Inhaltsanalysen von Fachgutachten haben sich wie gesagt weiter in die Innenwelt des Peer Review hineinbegeben. Ihre (oben dargestellten) methodischen Grenzen haben ebenfalls einen konzeptuellen Hintergrund: Es fehlt ihnen i.d.R. an einer Berücksichtigung des „inszenierten" Charakters (Neidhardt 1988: 85) solcher Texte. Gutachten werden (wie Publikationen) nur als stabile Dokumente, nicht aber als Ausfluss einer kommunikativen Praxis analysiert, die sich mit rhetorischen Mitteln in einem Interaktionszusammenhang positioniert. Hier fehlt es der Peer Review Forschung an einem Anschluss an die zahlreichen Studien zur Rhetorik der Wissenschaft wie sie insbesondere Wissenschaftshistoriker seit den 90er Jahren vorgelegt haben (etwa Bazerman 1988, Prelli 1989, Gross 1990, Myers 1990, Simons 1990, Dear 1991, Pera/Shea 1991, Locke 1992, Berkenkotter/Huckin 1995). Ein solcher Anschluss würde vor allem zwei Umstellungen verlangen:

Zum ersten muss die analytische Einheit des ‚Gutachtens' stärker kontextiert werden. Die Schriftfassung von Gutachten, ihr Dokumentcharakter, verführt ebenso wie der von Manuskripten dazu, sie als Gegenstände zu reifizieren. Ein Gutachten ist aber ebensowenig ein Gegenstand wie ein Manuskript. Es ist ein kommentierender Text, der in einer Dreiecksbeziehung von Autor, Gutachter und Herausgeber aufgespannt ist. Will man seine semioti-

von Wissenstypen: Die größten Studien bieten ein oberflächliches Wissen *über* den Peer Review, indem sie das Netz mathematischer Indikatoren darstellen, die die Verfahren als leicht messbare Eckpunkte abwerfen. Die Erfahrungsberichte und Argumente von Diskutanten stützen sich dagegen, meist unter Ermangelung empirischer Daten, auf ein Erfahrungswissen *aus dem* Peer Review. Das Problem besteht darin, dass das Wissen über den Peer Review nur einen Bruchteil dieses Wissens *im* Peer Review spiegelt: anekdotische Erfahrungen, rhetorische Skills, strategisches Know-How usw. Die Kluft zwischen dem Wissen der Teilnehmer und der Beobachter ist vermutlich das größte ‚Einfallstor' für die üble Nachrede, die zum Peer Review ebenso gehört wie die Beteuerungen seiner Unverzichtbarkeit.

[17] Neben der Frage der Rollendifferenzierung beim Konsensbedarf ist die nach der disziplinären Differenzierung wichtig, da der Peer Review in multiparadigmatischen Disziplinen (wie etwa der Soziologie) z. T. andere Formen annehmen kann als in anderen. Die Peer Review Forschung ist in dieser Hinsicht einerseits beeindruckend interdisziplinär, aber andererseits auch zu wenig disziplinär differenziert. Viele Autoren verfolgen hier in ihren Disziplinen eine Idee von Einheitswissenschaft, die die Wissenschaftssoziologie bereits gründlich erschüttert hat (Knorr Cetina 2002). Auch dieser Aufsatz kann nicht die angebrachte Differenzierung nach Fachkulturen leisten, er beschränkt sich darauf, sich an den Unterstellungen einer kohärenten Wissenschaft abzuarbeiten, in der viele Peer Review Forscher *ihren* Konsens suchen.

schen Funktionen verstehen, muss man ihn zum einen auf andere Textsorten im Peer Review beziehen: Herausgebervoten und Manuskripte, zum anderen muss man ihn viel stärker im Prozess der Anfertigung und Rezeption betrachten. Schon Bailar hatte (in der Diskussion mit Cicchetti 1991) in diesem Sinne moniert, dass es keine Studien darüber gibt, wie Herausgeber Gutachten eigentlich nutzen. Daran hat sich nichts geändert. Man weiß nichts darüber, wie Herausgeber die Beurteilungsprobleme lösen, die Gutachten genauso wie Manuskripte stellen. Man weiß auch nichts darüber, wie Gutachter die spätere Nutzung ihrer Texte durch Herausgeber und Autoren antizipieren und wie ihr Urteil durch Erwartungsverschränkungen – in Bezug auf seine Strenge oder seine schulenpolitische Tendenz – ‚voreingestellt' wird.

Zum zweiten müssen Gutachten statt als Dokumente von rationaler Kognition als *Sprechpraxis* ernstgenommen werden. Wissenschaft ist nicht einfach ein rationaler Prozess ‚im Kopf', sondern ein mit Argumenten geführter Aushandlungsprozess in der Kommunikation: Was Gutachter schreiben, ist nicht ‚was sie denken' (was immer das sein mag), sondern was sie – für bestimmte Adressaten, in bestimmten sprachlichen Formen – *mitteilen*. Auch in dieser Hinsicht hat die Peer Review Forschung einen gewissen szientistischen Bias, indem sie der Orientierungsleistung von Gütekriterien einen deutlichen Primat vor der Unsicherheit der Urteilsbildung gibt. Kriterien sind sicher Orientierungspunkte der individuellen Urteilsbildung, aber sie können deren Unsicherheit nicht beseitigen, weil ihre Zielkonflikte eine ‚Meinungsverschiedenheit mit sich selbst' erzeugen können. Ferner verpflichten Kriterien den Meinungsstreit sicher auf viele geteilte Standards, aber deren ‚Anwendung' kann ihn nicht schlichten, weil er sich wesentlich um ihre Rangordnung dreht. Daher sind ‚Kriterien' immer auch rhetorische Ressourcen, also strategisch ‚kontaminiert': Sie werden mobilisiert oder verschwiegen, mit Emphase versehen oder heruntergespielt.

Ein generelles Argument in dieser Hinsicht wurde von Seiten der Ethnomethodologie gemacht: Die Bedeutung von Kriterien oder Entscheidungsregeln kann nicht unabhängig von den situierten Praktiken verstanden werden, durch die sie implementiert werden (Garfinkel 1967). In einer konversationsanalytischen Studie zur Feststellung des Nachrichtenwerts in Redaktionskonferenzen von Zeitungen haben Clayman/Reisner (1998) dieses Argument ausgeführt: Kriterienlisten können das Spektrum der Aspekte beleuchten, die Herausgeber erwägen, aber sie haben einen geringen prognostischen Wert in Bezug auf Entscheidungen und sie gehen an den tatsächlichen Praktiken des *gatekeeping* vorbei, am Austausch von Redaktionskonferenzen, an der Evokation von Kriterien, an Vorschlägen, Aushandlungen etc. Diese situativen Praktiken sind keine Epiphänomene eines mentalen Beurteilungsprozesses, ‚*gatekeeping in action*' ist vielmehr essentiell eine öffentliche Sprechpraxis.

Dieses generelle (und wesentlich methodisch motivierte) Argument hat nun im Fall des Peer Review eine besondere theoretische Relevanz: Schriftliche Gutachten und die Kommunikation in Herausgebergremien unterscheiden sich vom ‚stillen Urteil' gewöhnlicher Leser gerade dadurch, dass sie in bestimmter Hinsicht selbst ‚publik' werden. Es handelt sich um halböffentliche Sprechhandlungen, die auf viel schärfere Weise in ihrer Verbindlichkeit kontrolliert werden als etwa das Geraune auf Tagungsfluren. Und die soziale Leistungsfähigkeit des Peer Review könnte u. a. eben darin begründet liegen, dass er Aspekte wissenschaftlicher Kommunikation ‚publiziert', die sonst ohne öffentliche Kontrolle keimen.

3.4 Kommunikationstheorie statt Publikationsfixierung

Die Quantifizierung von Zitationen zur Validierung von Peer Review Verfahren hatte den Vorzug, Gutachterurteile über Herausgeberentscheidungen hinaus zu kontextieren: in das Kommunikationsgeschehen disziplinärer Gemeinschaften. Gegen die vorliegenden Versuche sind aber zwei grundlegende Einwände zu machen:

Zum ersten fehlt es den Zitationsanalysen an der nötigen epistemologischen Skepsis gegen das Validierungsmotiv. Die Suche nach einem unabhängigen Maß wissenschaftlicher Güte hat wie die experimentellen Forschungsdesigns ein epistemologisches Problem mit der Historizität wissenschaftlicher Güteurteile. Ob man Zitationsquoten nun zur Verteidigung des Peer Review aufbietet oder lange Listen von ‚*rejected classics*' anfertigt (Gans/Shephard 1994, s. a. Crampanzano 1998a) – Aufsätzen prominenter Autoren, die zuerst abgelehnt, und nach Publikation bei einer anderen Zeitschrift zu vielzitierten Klassikern wurden – so handelt es sich in beiden Fällen um eine seltsame Besserwisserei der Spätgeborenen, die nicht damit rechnet, dass sie selbst jederzeit historisch überholt werden kann. Die Wissenschaftsforschung kann Validität nicht feststellen, weil es in der Wissenschaftspraxis selbst nach langer Zeit und auch bei anerkannt bahnbre-

chenden Publikationen noch erhebliche Divergenzen im Werturteil geben kann.[18]

Zum zweiten bestand ein wesentlicher methodischer Einwand gegen diese Untersuchungsdesigns darin, dass eine quantitative Zitationsmessung bei extrem geringen Fallzahlen wenig Sinn macht. Auch schon in der Diskussion um die Studie von Peters/Ceci war der Hinweis auf die Winzigkeit der Leserkreise ein wichtiges Argument. Er erklärte dort, warum Plagiate verkannt wurden und deckte die (autoritativ auftretende) Fiktion ‚des Feldes' oder ‚der Literatur' auf, die einem Gutachter bekannt sein soll. Man kann die Bedeutung des Arguments der kleinen Leserkreise kaum scharf genug formulieren. Sie liegt darin, dass die allermeisten Fachaufsätze in der Rezeption, den Zitaten und den Erinnerungen einer Disziplin völlig *spurlos* bleiben. Weder die Kritiker von Peters/Ceci, denen es um die Erklärung der Leistungsschwäche des Peer Review ging, noch die Zitationsanalytiker, denen es um den Nachweis seiner Leistungsstärke ging, lösen sich von einem tiefsitzenden Bias der Wissenschaftsforschung: der Überschätzung von Publikationen. Dieser Bias besteht aus zwei Komponenten: einer wissenspsychologischen (1.) und einer methodischen (2.)

1. Einerseits bilden Publikationen im Hinblick auf die Funktion der Reputationsverteilung eine Art Telos wissenschaftlicher Kommunikation i.S. von Stichweh (1994). Eine Publikation ist wissenspsychologisch ein ‚feierlicher Endpunkt', der dem Naturwissenschaftler Priorität zertifiziert, dem Geisteswissenschaftler zumindest den mentalen Abschluss einer intellektuellen Beschäftigung erlaubt. Andererseits handelt es sich soziologisch und historisch aber eher um ein lapidares Durchgangsstadium eines kollektiven Wissensprozesses. Ob man den Peer Review kritisiert, weil er ‚bahnbrechend innovative' Aufsätze verhindert, oder ob man ihn verteidigen will, weil er Aufsätzen ein oder zwei Zitate mehr im Jahr erwirtschaftet, – in jedem Fall überschätzt man drastisch das knappste Gut im Prozess wissenschaftlicher Kommunikation. Es sind nicht die Druckseiten, sondern die Aufmerksamkeit von Lesern. ‚*Publizität*' durch wissenschaftliche Aufsätze ist eine *Autorenfiktion*, die für die Mobilisierung der narzisstischen Brennstoffe wissenschaftlicher Arbeit nützlich ist, aber nicht zur Modellierung wissenschaftlicher Kommunikation taugt.

2. Einerseits ist die empirische Zugänglichkeit ihrer schriftlichen Selbstdokumentation (etwa für Zitationsanalysen) ein guter methodischer Grund für die Wertschätzung von Publikationen in der Wissenschaftskommunikation. Andererseits führen die an Publikationen (und Publikationsempfehlungen) anschließenden Quantifizierungschancen aber zu einem Zerrbild wissenschaftlicher Kommunikation, wenn der Erhebungsvorteil mit der Beschaffenheit des Gegenstands verwechselt wird. Gross (1990) meint, wenn man Wissenschaft wesentlich als Kommunikationprozess betrachtet, kann man sich nicht auf die Analyse der publizierten Endprodukte beschränken, man muss auch ganz andere Stadien der Metamorphosen von Laborprotokollen hin zu Lehrbüchern analysieren. Fröhlich (1998) stellt ein ähnliches Forschungsdefizit für die mündliche Kommunikation fest: Wenn die Teilnehmer an Kongressen regelmäßig die informelle Kommunikation (den Klatsch, die Tips, die Nachrede) wichtiger nehmen als die öffentlichen Vorträge, dann müssen die Sozialformen dieser Kommunikation vordringlich zum Gegenstand gemacht werden. Wissenschaftliche Kommunikation ist eben auch wesentlich mündliches Gerede oder ist, gerade im Fall des Peer Review, durch Diskretionsschranken vor Öffentlichkeit geschützt. Wenn sich die Peer Review Forschung aber vor allem an Publikationen orientiert, verfehlt sie von vornherein jene halb mündliche, halb schriftliche *informelle* Kommunikation, die ihren Gegenstand auszeichnet. Mit dem Peer Review müssen Aspekte wissenschaftlicher Kommunikation in den Blick genommen werden, die eben nicht publiziert werden. Und diese lassen sich nicht mit den theoretischen Prämissen der Publikationsforschung erfassen.

Wenn man den Impuls der Zitationsforschung aufgreifen will, Publikationsmomente in weiter gefasste Kommunikationsprozesse einzubetten, wird man daher erhebliche empirische und theoretische Anstrengungen brauchen, um das Verhältnis von Publikation und Kommunikation in der Wissenschaft zu klären. Was heißt eigentlich ‚Publikation' (außer ‚auf Dokumentation beruhender Zugänglichkeit'), wenn von ‚Publizität' angesichts kleinster Leserkreise kaum die Rede sein kann? Diese Frage kann im Rahmen dieses Aufsatzes nur aufgeworfen werden, es seien aber zumindest drei Aspekte genannt, auf die sich die Aufmerksamkeit richtet, wenn sich der Blick auf die Publikation lockert:

[18] Der heuristische Relativismus, den ich hier nahelege, ist übrigens grundsätzlich unabdingbar für eine professionell betriebene Wissenschaftsforschung. Nur wenn man auf ‚besserwisserische' Übergriffe verzichtet und wissenschaftliche Güteurteile konseqent bei den Forschern lässt, die sie zu einem bestimmten Zeitpunkt zu fällen haben, kann man darauf hoffen, das analytische Problem zu meistern, wissenschaftliche Praxis zu verstehen und zu erklären (und nicht zu bewerten).

1. *die Praktiken des Lesens*. Wenn man sich ein mitunter als Ersatz des Peer Review vorgeschlagenes laisser-faire-System vorstellt, das nur Autoren und Leser ohne den hierarchischen Filter der Begutachtung entscheiden lässt (Harnad 1998a), so wird die Rolle von Herausgebern und Gutachtern als Testleser und Vorkoster mit Orientierungsfunktion schnell deutlich. Der Peer Review ist vor allem eine *Einrichtung zur Kalibrierung der Lesezeit einer Disziplin* (Harnad 1998b: 9). Die Wissenschaftskommunikation muss Personen mobilisieren, die sich einer Verpflichtung zur Lektüre unterziehen und in ihren Gutachten – weit mehr als in Zitaten – Lektüre *dokumentieren* oder eben Dritte erkennen lassen, wo sie versäumt wurde oder am Verstehen scheiterte.

2. *die Funktion des Feedbacks*. Wenn man Publikationen nicht als Telos wissenschaftlicher Kommunikation betrachtet, sondern als Durchgangsstadium einer Raum und Zeit überspannenden Interaktion, dann sieht man besser, dass ein zur Lektüre und auf schriftliche Stellungnahme verpflichteter Leser Einwürfe in den Monolog des Autors macht, die für wissenschaftliche Kommunikation essentiell sind: nicht nur im Sinne der Optimierung kommunikativer Angebote, sondern auch i. S. der Erhöhung ihrer Responsivität (Harnad 1990). Zentral dürfte hierbei die Gutachterauswahl sein: ein Matching, das Personen in kommunikativen Kontakt bringt, die sich ‚viel zu sagen' haben, ohne sich zu kennen.

3. *das Verhältnis von Autor und Community.* Es gibt bekanntlich große Unterschiede zwischen Natur- und Sozialwissenschaften in der Stilisierung von Forschern als individuellen Autoren. Die Naturwissenschaften werden i.d.R. durch komplexe Arbeitsteilungen in der Laborarbeit und durch Koautorschaften bestimmt. Bakanic/McPhail/Simon (1987) legen nahe, dass *Peers* in den Sozialwissenschaften dagegen erst viel später in den Wissensprozess eingreifen: Hier findet die Kollektivierung der Wissensproduktion eben oft erst nach der Einreichung bei einer Fachzeitschrift statt.

Alle drei Aspekte haben eine gesteigerte Relevanz durch die technischen Innovationen der Wissenschaftskommunikation bekommen. Durch das Internet können Forscher sich in ihrem Publikationsgebaren finanziell unabhängig von Verlagen machen (wie etwa die Physiker mit dem Pre-Print-Archiv des Los Alamos National Laboratory), sie können billiger, schneller und adressatenspezifischer kommunizieren und sich von den Darstellungsbeschränkungen der Papierkommunikation befreien (zugunsten von Visualität, Akustik, Umfang). Darüber hinaus können sie sich auch eine größere „Freiheit" von den Prüf- und Selektionsverfahren des Peer Review versprechen. Banner (1988) sah in diesem Sinne in den Möglichkeiten des Internet starke Gründe für einen Bedeutungsverlust des Peer Review: Die elektronische Kommunikation bietet viele ‚niedrigschwellige' Publikationschancen bei schneller Distribution. Eine solche Freiheit kann bei den *Lesern* aber nur die „Informationsverdrossenheit" (Fröhlich 1994: 91) steigern, die Erleichterung des Publizierens verlangt gerade nach einer Verschärfung von Verfahren zur Redundanzreduzierung und zur Informationsbewertung. Das Internet lässt also die zentrale Funktion des Peer Review – die Steuerung von Leseraufmerksamkeit – unberührt: Peer Review Verfahren sind in ihrem Kern ‚medienunabhängig' (Harnad 1998a). Dennoch erlaubt das Internet natürlich Verfahrensvariationen, z. B. die erhebliche Beschleunigung von Begutachtungsprozessen, eine Verbesserung der Gutachterauswahl, die Erleichterung von Leserfeedback (die Überwindung der Stummheit des Papiers) – bis zum Übergang in neue Typen von Zeitschriften. Auch diese Phänomene wird man nicht mit einer auf Dokumente fixierten Publikationsforschung, sondern nur mit einer erweiterten Kommunikationsforschung analysieren können, die die sozialen Beziehungen zwischen Autoren, Gutachtern, Herausgebern und Lesern soziologisch rekonstruiert.

4. Eine Schlussbemerkung

Ich habe eingangs festgestellt, dass die Relevanz der Forschung zum Peer Review nicht nur in dessen wissenschaftlicher Bedeutung liegt, sondern auch in seiner Instrumentalisierung für politische Zwecke, und zwar sowohl professionsinterne als auch wissenschaftspolitische. Im Hinblick auf Professionspolitik war eine Debatte aufschlußreich, die im vergangenen Jahr in der Mitgliederzeitschrift der DGS um einen Artikel von Jürgen Gerhards (2002) geführt wurde. Sie verdient eine Anmerkung zur Sache und zur Form.

Zur Sache: Publikationserfolge in begutachteten Zeitschriften sind zweifellos ein besserer Indikator für das professionelle Standing eines Autors als reine Publikationsziffern (die freilich sehr wohl ‚Reputation' verschaffen können). Sie stehen eben nicht nur für Schreibfleiß und Schreiblust, sondern dafür, dass jemand mit seinen Argumenten einige systematisch prüfende Leser in einer Konkurrenz mit anderen Autoren überzeugen konnte. Ferner sind diese Publikationserfolge (anders als Günter Endruweit (2002) meint) oft auch ein besserer Indikator als Zi-

tationsziffern. Letztere sind wegen der normalerweise schwachen Fallzahl und der unklaren Valenz von Zitationen bloße Additionen einsilbiger Leserreaktionen, deren Bedeutung als Bewertung völlig unklar ist. Publikationserfolge müssen dagegen wegen der kontrollierten Qualität der gutachterlichen Lektüre, die ihnen notwendig vorausgeht, auch als *Rezeptionserfolge* gelten.

Zur Form: Wenn diese Erfolge in einer nonchalanten Quantifizierung in ein ‚Ranking' überführt werden, so steht dieser scheinbare ‚Indikationsoptimismus' für zwei Dinge: zum einen für die szientistische Sehnsucht, den komplexen sozialen Gehalt unabschließbarer Beurteilungsprozesse über neues Wissen in schlichten Maßzahlen abzuschneiden (die z. B. acht solide und zwei bahnbrechende Aufsätze einem flüchtigen Blick ‚vergleichbar' macht). Die aggregierten Ziffern brechen endlich aus dem sozialen Dickicht der Begutachtungen und den chronischen Unsicherheiten und Abwägungsproblemen ihrer Urteile aus. Zum anderen steht so ein Ranking, das hat Günter Burkart (2002) deutlich gemacht, für die Nutzung der Form eines wissenschaftssoziologischen Aufsatzes als professionspolitisches Instrument der Reputationszuweisung. Das so unschuldige „Warum nicht?" des einen und die Empörung der anderen stehen gleichermaßen für ein gegenüber dem Reputationsgedrängel unterentwickeltes *sachliches* Interesse an den Mechanismen und Folgen der Wissenschaftsevaluation.

Was die *externe* Evaluation von Forschung betrifft, so kann die Instrumentalisierung des Peer Review auf zwei Weisen problematische Effekte im Sinne einer Fehlsteuerung von Ressourcen haben. Zum einen auf Seiten der Rezeption von Gutachten: Außerhalb der Wissenschaft werden Gutachten tendenziell nicht mehr als Äußerungen-im-wissenschaftlichen-Meinungsstreit aufgefasst, sondern als autoritative Expertenäußerungen ‚der Wissenschaft' und diese Verkürzung gelingt umso eher, je geringer die Zahl der Gutachter (d. h. je geringer die Dissenschancen).

Zum anderen können solche Erwartungen der Politik auch entsprechende Sprecherpositionen *in* der Wissenschaft hervorbringen. Eben dies scheint das Gros der Peer Review Forschung wie auch der quantitativen Wissenschaftsevaluation zu bestätigen: Wenn etwa Dissens als *‚random'* gilt, übernimmt eine für Zwecke politischer Evaluation eingesetzte Wissenschaftsforschung ein Fremdstereotyp von Wissenschaft – dass diese sicheres und objektives Wissen generiere – in ihre Selbstbeschreibung. Diese bestätigt dann wiederum die Erwartungen (und Hoffnungen) der Politik, dass Wissenschaft politikferner sei als sie es tatsächlich ist; dass es in ihr nicht auch um Öffentlichkeit und das Gelingen von Kommunikation sowie um Diskursivität und Streit ginge: nämlich um Parteilichkeit und ihre Neutralisierung durch Verfahren, die Legitimität für hierarchiebedürftige Entscheidungen unter Professionskollegen beschaffen müssen. Aus dieser Spiegelung von Erwartungshaltungen führt nur eine reflexive Forschung zur Qualität von Verfahren der Qualitätsbeurteilung heraus, die nicht vorschnell eine vordergründige „Leistungsmessung" wissenschaftlicher Güte verspricht, sondern sich um die professionelle Evaluation eben der Instrumente bemüht, mit denen diese Leistungsmessungen erfolgen sollen.

Literatur

Abelson, P.H., 1980: Scientific Communication. Science 209: 60–62.

Amabile, T.M. / Glazebrook, A.H., 1982: A Negativity Bias in Interpersonal Evaluation. Journal of Experimental Social Psychology 18: 1–22.

Armstrong, J.S., 1997: Peer Review for Journals: Evidence on Quality Control, Fairness, and Innovation. Science and Engeneering Ethics 3: 63–84.

Bakanic, V. / McPhail, C. / Simon, R.J., 1987: The Manuscript Review and Decision-Making Process. American Sociological Review 52: 631–642.

Bakanic, V. / McPhail, C. / Simon, R.J., 1989: Mixed Messages: Referees' Comments on the Manuscripts They Review. The Sociological Quarterly 30: 639–654.

Banner, J.M., 1988: Preserving the integrity of peer review. Scholarly Publishing 19: 109–115.

Bazerman, C., 1988: Shaping Written Knowledge. The Genre and Activity of the Experimental Article in Science. Madison: University of Wisconsin Press.

Beck, U. / Hartmann, H., 1983: Wer ist der schönste im ganzen Land? Überlegungen zur Auswahl eines preiswürdigen Zeitschriftenaufsatzes. Soziale Welt 34: 257–269.

Berkenkotter, C. / Huckin, T., 1995: Genre Knowledge in Disciplinary Communication. Hillsdale, N.Y.: Lawrence Erlbaum Associates.

Blank, R.M., 1991: The Effects of Double-Blind versus Single-Blind Reviewing. American Economic Review 81: 1041–67.

Bös, D., 1998: Gedanken zum Refereesystem in ökonomischen wissenschaftlichen Zeitschriften. S. 47–72 in: F. Baltzarek / F. Butschek / G. Tichy (Hrsg.), Von der Theorie zur Wirtschaftspolitik – ein österreichischer Weg. Stuttgart: Lucius.

Bohn, C., 1999: Schriftlichkeit und Gesellschaft. Kommunikation und Sozialität in der Neuzeit. Opladen: Westdeutscher Verlag.

Bornstein, R.F., 1991: Manuscript Review in Psychology: Psychometrics, Demand Characteristics, and an Alter-

native Model. The Journal of Mind and Behavior 12 (4): 429–467.
Burkart, G., 2002: Die Faszination der Popsoziologie. Soziologie 3/2002: 47–52.
Campanario, J.M., 1998a: Peer Review for Journals as It Stands Today – Part I. Science Communication 19: 181–211.
Campanario, J.M., 1998a: Peer Review for Journals as It Stands Today – Part II. Science Communication 19: 277–306.
Ceci, S.J. / Peters, D., 1984: How Blind is Blind Review? American Psychologist 39: 1491–94.
Chubin, D. / Hackett, E., 1990: Peerless Science. Albany: Suny Press.
Cicchetti, Domenic V., 1991: The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences 14: 119–135. Diskussion: 135–186.
Cicchetti, D. / Eron, L., 1979: The Reliability of Manuscript Reviewing for the Journals of Abnormal Psychology, S. 596–600 in: Proceedings of the Social Statistics Section, Vol. 22, Washington: American Statistical Association.
Clayman, S. / Reisner, A., 1998: Gatekeeping in Action. Editorial Conferences and Assessments of Newsworthiness. American Sociological Review 63: 178–199.
Cole, S., 1983: The Hierarchy of the Sciences. American Journal of Sociology 89: 111–139.
Cole, S. / Rubin, L. / Cole, J.R., 1978: Peer Review in the National Science Foundation. Phase I of a Study. Washington: National Academy of Sciences.
Cole, J.R. / Cole, S., 1981: Peer Review in the National Science Foundation. Phase II of a Study. Washington: National Academy of Sciences.
Cole, S. / Cole, J. / Simon, G., 1981: Chance and Consensus in Peer Review. Science 214: 881–886.
Crane, D., 1967: The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals. The American Sociologist 2: 195–201.
Cronbach, L.J., 1981: Comment on „Chance and Consensus in Peer Review". Science 214: 1293.
Cyberconference 1999: Global Cyberconference on Peer Review in the Social Sciences, 28.5.–10.6.1999: www.sciencecity.org.uk/cyberconference.html
Daniel, H.-D., 1993: Guardians of Science. Fairness and Reliability of Peer Review. Weinheim: VCH.
Day, A., 1996: How to Get Research Published in Journals. Aldershot: Gower.
Day, R., 1998: How to Write and Publish a Scientific Paper. Cambridge University Press.
Dear, P. (ed.) 1991: The Literary Structure of Scientific Argument: Historical Studies. Philadelphia: University of Pennsylvania Press.
Endruweit, G., 2002: Wie misst man Reputation? Messtheoretische Überlegungen zu Jürgen Gerhards „Reputation in der deutschen Soziologie". Soziologie 4/2002: 33–41.
Epstein, W.M., 1990: Confirmational Response Bias Among Social Work Journals. Science, Technology, & Human Values 15: 9–38.
Fiske, D.W. / Fogg, L., 1990: But the Reviewers are Making Different Criticisms of my Paper! Diversity and Uniqueness in Reviewer Comments. American Psychologist 45: 591–598.
Forschungsförderung in Deutschland. Bericht der internationalen Kommission zur Systemevaluation der D.F.G. und der M.P.G., Hannover 1999.
Fröhlich, G., 1994: Der (Mehr-)Wert der Wissenschaftskommunikation. S. 84–95 in: W. Rauch et al. (Hrsg.), Mehrwert von Information. Konstanz: UVK.
Fröhlich, G., 1998: Optimale Informationsvorenthaltung als Strategem wiss. Komm. S. 535–549 in: H. Zimmermann / V. Schramm (Hrsg.), Knowledge Management und Kommunikationsysteme. Konstanz: UVK.
Fröhlich, G., 1999: Das Messen des leicht Messbaren. Output-Indikatoren, Impact-Maße: Artefakte der Szientometrie? S. 27–38 in: J. Becker / W. Göhring (Hrsg.), Kommunikation statt Markt. GMD-Report 61. Sankt Augustin.
Gans, J.S. / Shepherd, G.B., 1994: How are the Mighty Fallen: Rejected Classic Articles by Leading Economists. Journal of Economic Perspectives 8: 165–179.
Garfield, E., 1979: Citation Indexing. Its Theory and Application in Science, Technology, and Humanities. New York: Wiley & Sons.
Garfield, E., 1989: Citation Behavior - An Aid or a Hindrance to Information Retrieval? Current Contents 18 (1.5.): 3–8.
Garvey, W.D., 1979 (ed.): Communication. The Essence of Science. Pergamon Press.
Garvey, W.D. / Griffith, B.C., 1971: Scientific Communication. Its Role in the Conduct of Research and Creation of Knowledge. American Psychologist 26: 349–362.
Gerhards, J., 2002: Reputation in der deutschen Soziologie – zwei getrennte Welten. Soziologie 2/2002: 19–33.
Gross, A.G., 1990: The Rhetoric of Science. Cambridge: Harvard University Press.
Gross, A.G., 1990: Persuasion and peer review in science: Habermas's ideal speech situation applied. History of the human sciences 3: 195–209.
Hall, G.M. (Hrsg.), 1998: Publish or Perrish. Wie man einen wiss. Beitrag schreibt ohne die Leser zu langweilen oder die Daten zu verfälschen. Bern: Huber.
Hargens, L.L. / Herting, J.R., 1990a: A New Approach to Referees' Assessments of Manuscripts. Social Science Research 19. 1–16.
Hargens, L.L. / Herting, J.R., 1990b: Neglected Considerations in the Analysis of Agreement Among Journal Referees. Scientometrics 19: 91–106.
Harnad, S., 1982: Peer Commentary on Peer Review. A Case Study in Scientific Quality Control. Cambridge: Cambridge University Press.
Harnad, S., 1985: Rational Disagreement in Peer Review. Science, Technology, and Human Values 10: 55–62.
Harnad, S., 1990: Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. Current Contents 45: 9–13.
Harnad, S., 1995: Interactive Cognition. Exploring the Potential of Electronic Quote/Commenting. S. 397–414

in: B. Gorayska / J.L. Mey (Hrsg.), Cognitive Technology: In Search of a Human Interface. Amsterdam: Elsevier.

Harnad, S., 1998a: The Invisible Hand of Peer Review. Nature Web Matters. 5.11.1998.

Harnad, S., 1998b: Learned Inquiry and the Net: The Role of Peer Review, Peer Commentary and Copyright. Learned Publishing 4: 283–292.

Harnad, S., 1999: The Future of Scholarly Skywriting. S. 84–98 in: A. Scammell (Hrsg.), I in the Sky: Visions of the Information Future. NY Kretzenbacher: Aslib.

Hartmann, I. / Neidhardt, F., 1990: Peer Review at the Deutsche Forschungsgemeinschaft. Scientometrics 19 (5-6): 419–425.

He, A.W., 1993: Language Use in Peer Review Texts. Language in Society 22: 403–420.

Iser, W., 1972: Der implizite Leser. München: Fink.

Johnson, D.M., 1992: Compliments and Politeness in Peer-Review Texts. Applied Linguistics 13: 51–71.

Johnson, D.M. / Roen, D.H., 1992: Complimenting and Involvement in Peer Review. Gender Variation. Language in Society 21: 27–57.

Junge, M., 1993: Die Farben der Revue. Eine empirische Untersuchung der Buchbeurteilungen der Herausgeber der Soziologischen Revue. Soziologische Revue 16: 231–242.

Kalthoff, H., 1996: Das Zensurenpanoptikum. Eine ethnographische Studie zur schulischen Bewertungspraxis. Zeitschrift für Soziologie 25: 106–124.

Knorr Cetina, K., 2002: Wissenskulturen. Ein Vergleich naturwissenschaftlicher Wissensformen. Frankfurt: Suhrkamp.

Kretzenbacher, H.L. / Thurmair, M., 1992: Methoden des Textvergleichs zur Beschreibung wissenschaftlicher Textsorten – das Peer Review. In: K.-D. Baumann / H. Kalverkämper (Hrsg.), Kontrastive Fachsprachenforschung. Tübingen: Narr.

Lindsey, D., 1978: The Scientific Publication System in Social Science. San Fransisco: Jossey-Bass.

Lindsey, D., 1988: Assessing Precision in the Manuscript Review Process: A Little Better than a Dice Roll. Scientometrics 14: 75–82.

Lindsey, D., 1991: Precision in the manuscript review process: Hargens and Herting revisited. Scientometrics 22: 313–325.

Lock, S., 1985: A Difficult Balance: Editorial Peer Review in Medicine. London: Nuffield Provincial Hospitals Trust.

Locke, D., 1992: Science as Writing. New Haven: Yale University Press.

Mahoney, M.J., 1977: Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. Cognitive Therapy Research 1: 161–175.

Mahoney, M.J., 1987: Scientific Publication and Knowledge Politics. Journal of Social Behavior and Personality 2: 165–176.

McCartney, J.L., 1973. Manuscript Reviewing. Sociological Quarterly 14: 444–446.

McNutt, R.A. / Evans, A.T. / Fletcher R.H. / Fletcher S.W., 1990: The Effects of Blinding on the Quality of Peer Review. Journal of the American Medical Association 263 (10): 1371–1376.

Merton, R., 1985: Entwicklung und Wandel von Forschungsinteressen. Aufsätze zur Wissenschaftssoziologie. Frankfurt: Suhrkamp.

Myers, G., 1990: Writing Biology: Texts in the Social Construction of Scientific Knowledge. Madison: University of Wisconsin Press.

Neidhardt, F., 1986: Kollegialität und Kontrolle – Am Beispiel der Gutachter der Deutschen Forschungsgemeinschaft (DFG). Kölner Zeitschrift für Soziologie und Sozialpsychologie 38: 3–12.

Neidhardt, F., 1988: Selbststeuerung in der Forschungsförderung. Das Gutachterwesen der DFG. Opladen: Westdeutscher Verlag.

Pera, M. / Shea, W.(eds.), 1991: Persuading Science: The Art of Scientific Rhetoric. Canton, Mass.: Science History Publications.

Perlman, D. / Dean, E. 1987: The Wisdom of Salomon: Avoiding Bias in the Publication Review Process, S. 204–221 in: D.N. Jackson / J. Rushton (Hrsg.), Scientific Excellence. Origins and Assessment. Beverly Hills, London: Sage.

Peters, D. / Ceci, S., 1982: Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again". The Behavioral and Brain Sciences 5: 187–195. Reprint und Diskussion in Harnad 1982.

Powell, W.W., 1985: Getting Into Print. The Decision-Making Process in Scholarly Publishing. Chicago: University of Chicago Press.

Prelli, L., 1989: A Rhetoric of Science. Inventing Scientific Discourse. Columbia: University of South Carolina Press.

Roediger III, H.L., 1987: The Role of Journal Editors in the Scientific Process, S. 222–252 in: D.N. Jackson / J. Rushton (Hrsg.), Scientific Excellence. Origins and Assessment. Beverly Hills, London: Sage.

Root, L.S., 1987: Faculty Evaluation: Reliability of Peer Assessments of Research, Teaching, and Service. Research in Higher Education 26: 71–84.

Rosenblatt, A. / Kirk S.A., 1980: Recognition of Authors in Blind Review of Manuscripts. Journal of Social Service Research 3: 383–394.

Sahner, H., 1982: Zur Selektivität von Herausgebern: Eine Input-Output-Analyse der ‚Zeitschrift für Soziologie'. ZfS 11: 82–98.

Siegelman, S.S., 1991: Assassins and Zealots: Variations in Peer Review. Radiology 178: 637–642.

Simon, R.J. / Bakanic, V. / McPhail, C., 1986: Who Complains to Journal Editors and What Happens. Sociological Inquiry 56: 259–271.

Simons, H.W. (Hrsg.), 1990: The Rhetorical Turn. Invention and Persuasion in the Conduct of Inquiry. Chicago: University of Chicago Press.

Singer, B., 1989: The Criterial Crisis of the Academic World. Sociological Inqiry 59: 127–143.

Smigel, E.O. / Ross, H.L., 1970: Factors in the Editorial Decision. American Sociologist 5: 19–21.

Snizek, W.E. / Fuhrmann, E.R., 1979: Some Factors Affecting the Evaluative Content of Book Reviews in Sociology. American Sociologist 14: 108–114.

Sonnert, G., 1995: What Makes a Good Scientist? Determinants of Peer Evaluation among Biologists. Social Studies of Science 25: 35–55.

Spencer, N.J. / Hartnett, J. / Mahoney, J., 1986: Problems with Reviews in the Standard Editorial Practice. Journal of Social Behavior and Personality 1: 21–36.

Stichweh, R., 1994: Die Autopoiesis der Wissenschaft. S. 52–83 in: ders.: Wissenschaft, Universität, Professionen. Frankfurt: Suhrkamp.

Stossel, T.P., 1985: Refinement in Biomedical Communication: A Case Study. Science, Technology, & Human Values 10 (3): 39–43.

Thyer, B., 1994: Successful Publishing in Scholarly Journals. London: Sage (Bd. 11 der Survival Skills for Scholars Series).

Travis, G.D.L. / Collins, H.M., 1991: New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. Science, Technology, & Human Values 16 (3): 322–341.

Weingart, P., 2001: Die Stunde der Wahrheit? Zum Verhältnis der Wissenschaft zu Politik, Wirtschaft und Medien in der Wissensgesellschaft. Weilerswist: Velbrück.

Wilson, J.D., 1978: Peer Review and Publication. Journal of Clinical Investigation 61: 1697–1701.

Whitehurst, G.J., 1984: Interrater Agreement for Journal Manuscript Reviews. American Psychologist 39: 22–28.

Zimmermann, K., 2000: Spiele mit der Macht in der Wissenschaft. Passfähigkeit und Geschlecht als Kriterien für Berufungen. Berlin: Edition Sigma.

Zuckerman, H. / Merton, R., 1971: Patterns of Evaluation in Science. Minerva 9: 66–100. Deutsche Fassung: Institutionalisierte Bewertungsstrukturen in der Wissenschaft. S. 172–216 in Merton 1985.

Summary: Peer review research is the sphere of science studies dealing with the central mechanism of the self-evaluation of scholarly practice. This article offers a critical review of this research. It seems that peer review research is deeply involved in the selfsame evaluative process which it should be observing with professional distance. The reason is a lack of sociological conceptualization of this activity. The article pleads for a theoretical reorientation from persons to social processes, from measures of reliability to acknowledgement of dissent, from cognition to speech and writing practices, and from publication counts to research in communication. Peer review is not a scientific measurement of the quality of publications, but a social institution for the calibration of reading time within a discipline.