Sonderdruck aus:
Hildegard Matthies · Dagmar Simon (Hrsg.)
Wissenschaft unter Beobachtung
2008, 357 Seiten, Broschur, € 39,90
VS Verlag für Sozialwissenschaften, Wiesbaden

ISBN 978-3-531-15457-2

Hildegard Matthies · Dagmar Simon (Hrsg.)

Wissenschaft unter Beobachtung

Effekte und Defekte von Evaluationen

Sonderdruck

Bruno S. Frey[1]

Evaluitis – eine neue Krankheit

Einleitung

In den letzten Jahren ist eine neue Krankheit ausgebrochen, die sich fieberhaft ausbreitet: Alles und jedes wird unablässig evaluiert. Unter „Evaluation" wird hier eine nachträgliche Einschätzung der Leistung einer Organisation oder Person durch von außen kommende Experten verstanden.[2] Der vorliegende Beitrag konzentriert sich auf Evaluationen im staatlichen Auftrag, deren vornehmliches Ziel es ist, die geeignete Zuteilung finanzieller Mittel zu unterstützen.

Von der Krankheit ist ganz besonders die Wissenschaft befallen. In immer kürzeren Abständen werden ganze Universitäten, Fakultäten, Fachbereiche, Institute, Forschungsgruppen und einzelne Forschende begutachtet. Evaluationen und daraus abgeleitete Rankings sind heute Allgemeingut geworden. Evaluationen dienen häufig der legalen und bürokratischen Legitimation staatlich finanzierter Universitäten, die staatlich verordneten Regeln unterworfen sind (Knorr Cetina 2006: 11). Entsprechend wird von einer „audit explosion" (Power 1994), einer „audit society" mit ihren „rituals of verification" (Power 1997), vom „age of inspection" (Day und Klein 1990) oder vom „evaluative state" (Neave 1988) gesprochen.

Im Folgenden soll auf einige wenig diskutierte, verborgene und gewöhnlich vernachlässigte Kosten von Evaluationen aufmerksam gemacht werden. Bei der Entscheidung, ob eine Evaluation durchzuführen ist – sofern darüber überhaupt noch entschieden wird –, bleiben diese Kosten in der Regel unberücksichtigt. Das hat zur Folge, dass der Nettonutzen dieses Instrumentes *systematisch* überschätzt wird und sowohl *Anzahl* als auch *Intensität* der durchgeführten Evaluationen höher sind als gesellschaftlich sinnvoll wäre. Insofern lässt sich auch von einer Krankheit namens „Evaluitis" sprechen. Damit soll jedoch keineswegs ein Argument gegen Evaluationen *an sich* vorgebracht werden; in manchen Fällen erweisen sich diese durchaus als notwendig und sinnvoll. Das Argument lautet allerdings auch nicht, die heutigen Evaluationen seien nur mangelhaft und ließen sich ohne weiteres verbessern. Die hier vorge-

1 Ich bedanke mich für wertvolle Hinweise bei Margit Osterloh, Reiner Eichenberger und Simon Lüchinger. Einige Einsichten verdanke ich meiner eigenen Tätigkeit als Evaluator verschiedener Forschungseinrichtungen in unterschiedlichen Ländern.

2 Diese Definition entspricht sowohl dem Alltags- als auch dem Wissenschaftsverständnis; vgl. z. B. Brook (2002: 173): "By evaluation, I shall mean the situation where visiting experts come from outside your organization or system and say what they think about it."

brachten Einwände sind vielmehr grundsätzlicher Natur und lassen sich nicht einfach beiseite räumen, indem die Evaluationen differenzierter werden.[3] Denn verbesserte – und das heißt intensivere – Evaluationen können möglicherweise noch zu einer Verschärfung der aufgeführten fundamentalen Probleme führen.

Zum Thema Evaluation liegt eine große Fülle von Literatur vor, deren Erkenntnisse hier nicht im Einzelnen wiederholt werden sollen.[4] In diesem Beitrag wird somit nicht auf die sattsam bekannten Kosten in Form von Material und Zeit auf Seiten der Evaluierenden und der Evaluierten eingegangen.[5] Ebenso wenig wird der offensichtliche Nutzen von Evaluationen für die Entscheidungsbildung diskutiert. Im Zentrum der Überlegungen stehen gerade solche Aspekte, die bei der Anwendung in der Praxis gewöhnlich vernachlässigt werden.[6] Betont wird, dass es *valable Alternativen zu Evaluationen* gibt. Diese Vorstellung widerspricht einer häufig geäußerten Meinung, Evaluationen seien absolut notwendig, weil ansonsten reine Willkür herrschen würde.[7] Zwar stimmt es, dass auf Evaluationen als Instrument der Qualitätssicherung nicht vollständig verzichtet werden kann. Aber, so lautet die hier vertretene These, sie ließen sich weitgehend reduzieren, vorausgesetzt, es werden mittels geeigneter Institutionen angemessene Anreize zur Leistungsverbesserung vermittelt und das Schwergewicht wird auf eine sorgfältige *vorherige* Auswahl von Personen gelegt.

Im ersten Teil des Beitrags werden die vernachlässigten Kosten von Evaluationen diskutiert. Kapitel 1 befasst sich mit der durch Evaluationen verursachten Anreizverzerrung bei den Evaluierten, Kapitel 2 mit der induzierten Verkrustung und Kapitel 3 mit dem verfehlten Entscheidungsansatz und damit dem geringen Nutzen für die Entscheidungsbildung. Im zweiten Teil (Kapitel 4 und 5) werden die Alternativen zu Evaluationen behandelt. Es wird das Argument entwickelt, dass ein gewünschtes Verhalten auch mittels institutioneller Änderungen und einer sorgfältigen Personenauslese erzielt werden kann. Im letzten Teil (Kapitel 6) werden abschließende Überlegungen ausgestellt.

---

3 Wobei es sicherlich eine sinnvolle Verbesserung von Evaluationen gibt, vgl. etwa die Vorschläge für ein stärker diskursiv ausgerichtetes und flexibles Vorgehen von Chapman (2006) und Knorr Cetina (2006) im Sonderheft *Accountability in Research* der Zeitschrift *Foresight Europe*.

4 Vgl. z. B. Broadfoot (1996), Backes-Gellner/Moog (2004), De Bruijn (2002), Max-Planck-Gesellschaft (2002), Russon/Russon (2000), Stockmann (2004), sowie einschlägige Zeitschriften wie etwa *Evaluation, Evaluation Review* oder das *American Journal of Evaluation*. Speziell zu Evaluationen in der Wissenschaft vgl. Bräuninger/Haukap (2003), Cash/Clark (2001), Daniel/Fisch (1988), Daniel (1993), Jordan (1989), Klostermeier (1994), Kozar (1999), Röbbecke/Simon (1999, 2001) und die Zeitschriften *Research Evaluation* und *Scientometrics*.

5 Ein Zitat aus dem *Economist* (2002: 69) soll genügen: Die amerikanischen Business Schools beklagen sich über "the huge amount of staff time involved in replying to polisters' questions".

6 Diese Aspekte werden zwar in der Literatur durchaus gelegentlich erwähnt, aber in der Praxis kaum oder gar nicht beachtet. Siehe etwa die frühe Analyse bei Ridgway (1956: 240), der davon spricht, dass "the cure is sometimes worse than the disease".

7 So etwa bei Holcombe (2004), Royal Netherlands Academy of Arts and Sciences (2005), Starbuck (2004), Weingart (2005).

1 *Evaluation verzerrt Anreize*

Das Instrument der Evaluation verändert das Verhalten der davon betroffenen Personen in systematischer, aber auch unbeabsichtigter Weise. Es darf somit nicht davon ausgegangen werden, dass Individuen (und entsprechend Institutionen) infolge einer Evaluation ihr Verhalten in der von den Evaluierenden gewünschten Weise verändern, das heißt zielorientierter und effizienter arbeiten. Eher werden durch Evaluationen unerwünschte Verzerrungen im Verhalten ausgelöst: (A) eine Konzentration auf das, was gemessen wird; (B) eine Verdrängung intrinsischer Arbeitsanreize, wodurch vor allem die Originalität Schaden nimmt; und (C) eine Manipulation der Kennziffern.

1.1 Was nicht gemessen wird, zählt nicht (mehr)

Das Phänomen des Multitasking wird in der Wirtschaftswissenschaft seit einigen Jahren intensiv diskutiert[8]: Die Vorgesetzten (Prinzipale) legen die Maßstäbe fest, mit denen die Leistung einer Institution oder einer Person beurteilt wird. Es gibt jedoch keine Tätigkeit – außer möglicherweise der einfachsten Fließbandtätigkeit –, für die sich *alle* relevanten Aspekte definieren und messen ließen. Daher neigen die zu beurteilenden Personen dazu – oder werden sogar gezwungen –, sich bei ihrer Arbeit ausschließlich auf die gemessenen Kriterien zu konzentrieren und alles andere beiseite zu lassen. In den vielen Fällen, in denen nur die Inputs erfasst werden, ist die Verzerrung besonders gewichtig, weil dann die Produktivität völlig vernachlässigt wird.

In der Wissenschaft hat das Multitasking-Problem besonders starke Auswirkungen. Viele Universitäten sind dazu übergegangen, einfach die Zahl von Publikationen eines Forschenden zu zählen. Nur wenn eine bestimmte Zahl pro Jahr überschritten wird, steht der Weg für eine Professur oder eine Beförderung offen. "The result is a well-documented tendency to produce large numbers of articles based on trivial research results that are easily published" (Tucci 2006: 28). Ein weiteres Leistungskriterium, das heute fast überall angewendet wird, ist die Anwerbung von Drittmitteln (so etwa bei den Forschungseinrichtungen der Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz und der Max-Planck-Gesellschaft). Dass sich mit diesem Kriterium weder der Sinn noch die Produktivität der damit finanzierten Forschung erfassen lassen, dürfte augenscheinlich sein. Seine Verbreitung verdankt sich allein der Tatsache, dass Geldströme besonders leicht messbar sind. Wenn jedoch eine wissenschaftliche Einheit nach diesem Kriterium beurteilt wird, ist sie gezwungen, sich um Drittmittel zu bemühen und dafür weniger gut messbare Forschungs- und Lehrleistungen zu vernachlässigen. Selbst die Messung von Forschungsleistung mittels Zitierungen – was wesentlich näher beim gewünschten Output liegt – führt zu Verzerrungen. So bemerkt etwa Lindsay: "Citation counts as a measure of quality may often be measur-

8 Vgl. z. B. Daily/Dalton/Cannella (2003), Gibbons (1998), Holmstrom/Milgrom (1991), Suvorov/van de Ven (2006).

ing what is measurable rather than what is valid" (Lindsay 1989: 200). Vernachlässigt wird dabei die Übertragung wissenschaftlicher Erkenntnisse in die Praxis mittels Publikationen in populären Organen, allgemeinbildender Vorträge, Beratungstätigkeit sowie die universitäre Selbstverwaltung und die gesamte Lehrtätigkeit. Diese Probleme sind zwar wohlbekannt (vgl. z. B. Daniel 1993), aber es werden häufig daraus die falschen Schlüsse gezogen. Statt weniger Gewicht auf solche Evaluationen zu legen, wird versucht, die bislang vernachlässigten Aspekte auch noch quantitativ zu erfassen. Dies wird jedoch *nie* im vollen Umfang möglich sein. Das Multitasking-Problem wird deshalb auf immer schwerer messbare Aspekte verlagert, ohne dass dadurch die Verzerrung der Anreize beseitigt würde. Vielmehr kommt es zu einem dauernden Wettlauf zwischen den Evaluierten und den Evaluierenden. Das Ergebnis sind immer aufwändigere Evaluationsprozesse, die den Forschenden immer weniger Zeit für ihre eigentlichen Tätigkeit lassen: "Success in the evaluation process can become a more significant target than success in research itself" (Brook 2002: 176).

Dass die „Optimierung" des Evaluationsprozesses keine Lösung für die angesprochene Problematik bietet, lässt sich auch daran erkennen, dass selbst eine vollständige Erfassung von Zitierungen zu Verzerrungen im Verhalten führen würde. Sobald die Forschenden wissen, dass ihre wissenschaftliche Leistung an diesem Kriterium gemessen wird, werden sie veranlasst, sich Forschungsfragen zuzuwenden, die der augenblicklichen Mode entsprechen und daher häufige Zitierung gewährleisten. In vielen Disziplinen dürfte damit die angewandte Forschung benachteiligt werden.

1.2 Verdrängung intrinsischer Arbeitsanreize

Die mit der Evaluation einhergehende Messung und Beurteilung der Leistung beeinflusst die Arbeitsmotivation negativ, weil eine solche Bewertung von den Betroffenen in der Regel als *kontrollierend* empfunden wird. Dieser Effekt ist in der Sozialpsychologie in Hunderten von Laborexperimenten analysiert worden (vgl. die umfassende Metastudie von Deci/Koestner/Ryan 1999 sowie Cameron/Banko/Pierce 2001) und in der Ökonomik unter der Bezeichnung „Verdrängungseffekt" (Bénabou/Tirole 2003; Fehr/Gächter 2002; B. Frey 1992, 1997) empirisch anhand von Felduntersuchungen bestätigt worden (eine Übersicht geben B. Frey/Jegen 2001). Der Verdrängungseffekt besagt, dass infolge der als kontrollierend empfundenen Evaluation die intrinsische Arbeitsmotivation abnimmt, während die extrinsisch bestimmten Anreize an Gewicht gewinnen. Dabei vermindert sich die Gesamtleistung nicht notwendigerweise, sondern steigt sogar für manche Evaluierte, wie das britische „Research Assessment Exercise" festgestellt hat. Gemäß Brook (2002: 176) "[...] we can safely say that the average activity has increased" – zumindest in der von der Evaluation erfassten Dimension. Es darf jedoch bezweifelt werden, ob die Auswirkungen auch für die Qualität und Originalität der Forschung günstig waren. Denn wie Amabile (1996, 1998) gezeigt hat, ist die intrinsische Motivation für innovative wissenschaftliche Arbeit von entscheidender Bedeutung. Hinzu kommt, dass es gerade die bahnbrechende

Forschung ist, die Gefahr läuft, gering geschätzt zu werden, weil sie gegen den Konsens der Evaluierenden verstößt. Historische Untersuchungen (Fischer 1998; Gillies 2006) zeigen, dass viele besonders wichtige Forschungsergebnisse dem jeweiligen Zeitgeist (im Sinne der „normal science" von Kuhn 1962) widersprachen – was auch heißt, dass sie in einer Evaluation schlecht beurteilt worden wären.

Allerdings wird die intrinsische Forschungsmotivation nicht zwangsläufig durch eine Evaluation verdrängt, sondern kann sich sogar steigern, wenn die Betroffenen die Evaluation als unterstützend erleben (vgl. Heckhausen 1989). Das gleiche gilt, wenn die Evaluierten die ihnen zukommende Aufmerksamkeit genießen und sich kurzfristig mehr anstrengen (Hawthorne-Effekt). Beide Bedingungen dürften zutreffen, wenn die Evaluation neu eingeführt wird. Je mehr sie jedoch zu einer unablässigen Übung wird, umso stärker wird sie als kontrollierend empfunden, und die intrinsische Forschungsmotivation wird zunehmend verdrängt. Daraus folgt, dass extrinsische Anreize innovative Forschung tendenziell zerstören; in diesem Falle wirken sie dysfunktional (vgl. auch Kogut 2006: 4).

Der Verdrängungseffekt ist quantitativ schwer zu fassen, weswegen er leicht vernachlässigt wird. Man kann jedoch davon ausgehen, dass in dem Umfang, in dem dies der Fall ist, zu viel, zu häufig und zu intensiv evaluiert wird.

1.3 Manipulation der Leistungskriterien

Wenn ein Indikator für die eigene Position wichtig wird, wird ein starker Anreiz ausgeübt, diesen Indikator zum eigenen Nutzen zu beeinflussen. Dieser allgemeine Zusammenhang ist in der Volkswirtschaftslehre als „Goodhart's Law"(1975) oder „Lucas Critique" (1976) bekannt und empirisch auf der Makroebene gut nachgewiesen (vgl. z. B. Brück/Stephan 2006; Chrystal/Mizen 2003). Er gilt auch auf der Mikroebene. So können Schulleitungen etwa ihre Beurteilung beeinflussen, indem sie die Schüler auf die Examensaufgaben hin trainieren („teaching to the test") oder schlechte Schüler unter irgendwelchen Vorwänden von den entsprechenden Tests ausschließen, um die Ergebnisse ihrer Schule künstlich zu verbessern (zur empirischen Evidenz für die Vereinigten Staaten vgl. Figlio/Getzler 2003). Oder Manager beeinflussen die Leistungskriterien, sobald ihr Einkommen davon abhängig ist. Sie treiben zum Beispiel die Aktienpreise (kurzfristig) in die Höhe, wenn ein Teil ihres Gehaltes in Form von Aktienoptionen ausgezahlt wird (vgl. Osterloh/B. Frey 2005; B. Frey/Osterloh 2000a, 2000b, 2005).

Manipulationen dieser Art haben sich auch in der Wissenschaft verbreitet, seit die Forschungsleistung im Zuge von Evaluationen an der Zahl der Publikationen und Zitierungen gemessen wird. Die Forschenden haben schnell gelernt, wie sich die gemessene Forschungsleistung beeinflussen lässt: "People learn how to manage the reporting of performance" (Chapman 2006: 13). So werden etwa Wissenschaftler mit entsprechenden Leistungsausweisen vorübergehend an eine Universität verpflichtet, um dieser zu einem guten Abschneiden bei einer Evaluation zu verhelfen. Nicht sel-

ten haben diese Forschenden nur eine lose oder gar keine Beziehung zu den evaluierten Universitäten oder ihre Forschungsleistung wird von mehreren Universitäten gleichzeitig benutzt. Für die Wissenschaftskultur schädlicher ist das Hochjubeln von Ergebnissen in der Forschung weit über deren Bedeutung hinaus. So herrscht etwa ein verstärkter Anreiz, nur noch erfolgreiche Tests zu publizieren und die negativen Ergebnisse zu verschweigen oder sogar zu beseitigen. Noch verheerender ist der Anreiz zum Betrug mittels Fälschung von Forschungsergebnissen. In einem Experiment wurde gezeigt, dass Personen, die sich kontrolliert fühlen, weit eher zu betrügen bereit sind als solche, die sich nicht kontrolliert fühlen (Schulze/Frank 2003). Dass dieses Verhalten auch in der Realität des Wissenschaftsbetriebs vorkommt, haben verschiedene Skandale in der letzten Zeit bewiesen (vgl. z. B. Bedeian 2003; B. Frey 2003; McCabe; Trevino/Butterfield 1996).

Die durch Evaluationen verursachte Vernachlässigung nicht (einfach) messbarer Forschungsleistungen, Verdrängung intrinsischer Arbeitsmotivation und Manipulation der Leistungskriterien führt zu einem paradoxen Ergebnis. Der Evaluationsprozess induziert gerade jenes dysfunktionale Verhalten, das die Evaluation zu verhindern sucht. Anschauliche Beispiele dafür finden sich in einer Untersuchung von Business Schools (Hopwood 2005), in der nachzulesen ist, dass jüngere Forschende gerne und oft ihre letzten Veröffentlichungen in den führenden wissenschaftlichen Zeitschriften verkünden, aber nur selten darüber sprechen, welche Ideen in diesen Veröffentlichungen enthalten sind:

> "This [...] attitude [...] has now spread over the whole world, with researchers discussing how many times their publications got a 'hit' in the top journals without ever revealing (because they think it is unimportant) what the subject of their research actually was. [...] Scientists are no longer seen as defenders of truth, but more as defenders of their own interests in 'media driven' (or publicity-driven) science." (Tucci 2006: 27)

Dieses Verhalten ist auch auf wissenschaftlichen Konferenzen im Bereich der Volkswirtschaftslehre zu beobachten. Auf den Tagungen des *Vereins für Socialpolitik*, der *European Economic Association*, der *American Economic Association* und der *International Economic Association* (um nur einige zu nennen) gibt es gerade unter den jüngeren Teilnehmenden im kleinen Kreis kaum mehr eine Diskussion über inhaltliche Aspekte ihrer Forschung. Schon nach kurzer Zeit dreht sich das Gespräch nur noch um die Publikationserfahrungen, die man gemacht hat, und wie man am erfolgreichsten veröffentlichen kann.

2 *Induzierte Verkrustung*

Evaluationen bewirken „Lock-in-Effekte" sowohl (A) auf Seiten der Evaluierten als auch (B) auf Seiten der Evaluierenden. Wenn sich die Bedingungen ändern, insbesondere wenn sich herausstellt, dass Evaluationen weniger erfolgreich sind als bisher

angenommen, verhindern starke Kräfte, dass die Häufigkeit und Intensität der Evaluationen vermindert wird.

2.1 Die Situation der Evaluierten

Die Angehörigen einer Institution oder einzelne Forschende, für die eine Evaluation vorgesehen ist, können sich nicht gegen deren Durchführung zur Wehr setzen. Dies gilt selbst dann, wenn sie gute Gründe vorbringen können, dass sich eine bestimmte Evaluation für ihre Verhältnisse nicht eignet, zum Beispiel weil sich ein allzu großer Teil der Leistungen einer Bewertung und Messung entzieht. Man kann ihnen leicht entgegenhalten, sie hätten nur Angst vor dem Ergebnis der Evaluation. Da eine Evaluation typischerweise mit einer Mittelvergabe einhergeht, müssen sie sich wider besseren Wissens an dem Verfahren beteiligen. Sie tun sogar gut daran, begeistert mitzumachen. Nach außen wird dadurch der Anschein erweckt, die Evaluierten seien von den Vorzügen einer Evaluation überzeugt, also ein Einverständnis vorgetäuscht, das in Wirklichkeit nicht vorhanden ist. Damit wird einer zynischen Haltung zur Wissenschaft und deren Ergebnissen Vorschub geleistet.

2.2 Die Situation der Evaluierenden

Die Institutionen und Personen, welche die Evaluation durchführen, haben ein direktes Einkommens- und Karriereinteresse. Besonders ausgeprägt ist dieses Interesse bei privaten Anbietern, aber auch bei staatlichen Institutionen, deren Bedeutung und Budgetzuweisungen vom Fortbestand ihrer Evaluierungstätigkeit abhängt. Sie sind deshalb bestrebt, Evaluationen auf immer weitere Bereiche auszudehnen, zu intensivieren und in immer kürzeren Abständen durchzuführen. Am vorteilhaftesten für sie ist eine kontinuierliche Evaluation, wofür sich viele Argumente vorbringen lassen. Hingegen werden die negativen Aspekte von Evaluationen, wie etwa ihre im letzten Abschnitt aufgeführten verborgenen Kosten, heruntergespielt. Dieses aktive Lobbying oder „rent seeking" trägt zur Ausweitung der Evaluationen bei. Zugleich wird den Alternativen zur Evaluation – die in den Kapiteln 4 und 5 genannt werden – wenig oder gar kein Raum gegeben.

3 Geringer Nutzen von Evaluationen für Entscheidungen

Es gilt als selbstverständlich, dass die durch eine Evaluation gewonnenen Informationen wesentlich dazu beitragen, die Entscheidungen über die Planung und Förderung wissenschaftlicher Forschung zu verbessern. Es fällt schwer zu sehen, warum diese

zusätzlichen Informationen nicht so nützlich sind, wie sie auf den ersten Blick erscheinen. Dafür gibt es vor allem zwei Gründe.

3.1 Geringer Informationsgewinn

In den Scientific Communities ist häufig auch ohne Evaluationen sehr wohl bekannt, welche Institutionen und Personen gute Forschung betreiben. Bestätigt die Evaluation diese Annahmen, ist wenig oder nichts gewonnen. Kommt sie hingegen zum gegenteiligen Ergebnis, wird dieses zu Recht angezweifelt. Das gleiche gilt natürlich auch umgekehrt, wenn sich bei der Evaluation ein gutes Ergebnis für eine Institution herausstellt, die in der „Gelehrtenrepublik" einen schlechten Ruf hat. Es wird deshalb in beiden Fällen schwer fallen, die Ergebnisse der Evaluation politisch zum Tragen zu bringen.

Der Widerstand gegen die Ergebnisse einer Evaluation ist mit Sicherheit asymmetrisch. Wer gut eingeschätzt wird, ist erfreut und hofft auf höhere Budgetzuweisungen. Wer hingegen schlecht eingeschätzt wird, wird große Anstrengungen unternehmen, sich gegen die Auswirkungen zu wehren. Wie im folgenden Abschnitt gezeigt wird, stehen dafür gute Argumente zur Verfügung. Auf jeden Fall kann nicht davon ausgegangen werden, dass negative Evaluationsergebnisse die gewünschten Wirkungen erzeugen. Oft sind sie nur symbolischer Natur.

Nur in den – seltenen – Fällen, in denen in den Scientific Communities keine Einigkeit über die Qualität eines Forschenden oder einer Forschungsorganisation herrscht, kann eine Evaluation hilfreiche Informationen liefern. Allerdings dürften staatliche Evaluationen und Entscheidungen nur im Ausnahmefall überraschenden Ergebnisse liefern, vielmehr werden sie in der Regel mehr oder weniger die durchschnittliche Einschätzung bestätigen. Dies bedeutet wohl auch, dass die Mittelzuweisung zumindest im Vergleich zu anderen Institutionen wenig verändert wird. Der (hohe) Evaluationsaufwand lohnt sich deshalb nicht zwingend.

Zudem stellt sich die Frage, für wen die produzierten Informationen eigentlich gedacht sind. Es gibt viele Auftraggeber von Evaluationen, die für sich in Anspruch nehmen, das gesellschaftliche Interesse zu vertreten. Wer aber definiert die Standards? Das Problem stellt sich besonders im Falle von Forschungsprojekten, die – was häufig der Fall ist – aus unterschiedlichen Quellen finanziert werden, etwa durch die Europäische Union, die nationale Regierung, eine untergeordnete staatliche Verwaltungseinheit, eine der vielen Forschungsstiftungen oder die eigene Forschungseinrichtung. Denn der einzelne Forschende richtet sich im Allgemeinen nach den Anforderungen der internen Finanzierungsquelle (vgl. Tucci 2006: 29), die ja auch bestimmend für seine Karriere ist.

3.2 Für Entscheidungen irrelevante Information

Evaluationen suchen das bestehende Leistungsniveau anhand einer großen Zahl von Indikatoren wie etwa Publikations- und Zitierhäufigkeit oder Lehrerfolg zu erfassen. Für politische Entscheidungen sind diese Informationen jedoch von geringer Bedeutung, denn es bleibt zunächst völlig offen, was daraus zu schließen sei. Sollten den für schlecht befundenen Institutionen und Forschenden die Mittel gekürzt werden? Oder sollte man ihnen nicht gerade zusätzliche Mittel bewilligen, damit sie sich verbessern können? Empfehlen sich dann nicht auch Mittelkürzungen für die als gut bewerteten Institutionen und Forschenden, die ja ohnehin schon erfolgreich sind? Diese Fragen lassen klar erkennen, dass in der politischen Auseinandersetzung um Mittelzuweisungen noch alles völlig offen ist.

Im Idealfall sollte eine Evaluation den marginalen Effekt einer Änderung der Mittel erfassen: Was würde geschehen, wenn einer Institution oder einem Forscher mehr (oder weniger) Mittel zur Verfügung stünden? Diese Frage ist allerdings äußerst schwierig zu beantworten, weil zahlreiche Bedingungen zu berücksichtigen sind. Eine auf marginale Änderungen abstellende Evaluation ist wesentlich aufwändiger als die heute üblichen Ansätze, was sich negativ auf das Verhältnis von Kosten und Nutzen der Evaluationen auswirkt. Außerdem bleibt auch bei diesen Evaluationen offen, wie die Ergebnisse in der politischen Auseinandersetzung aufgenommen würden. Aus diesem Grund ist es ratsam, sich ernsthaft mit den Alternativen zu Evaluationen zu beschäftigen.

4 Institutionelle Alternativen

Die Art und Weise, wie eine Institution konstruiert ist, vermittelt bestimmte Anreize und beeinflusst damit systematisch das Verhalten von Personen. Dies ist die grundlegende Botschaft der modernen Ökonomik (vgl. B. Frey 1990, 2001; Kirchgässner 2000), insbesondere der „Institutionellen Ökonomik" (z. B. Erlei/Leschke/Sauerland 1999; Richter/Furubotn 1999) sowie der „Theorie der Wirtschaftspolitik" (B. Frey/Kirchgässner 2002). Diese empirisch in Hunderten von Studien nachgewiesenen Wirkungen brauchen an dieser Stelle nicht weiter ausgeführt zu werden. Aber an einem konkreten Beispiel soll gezeigt werden, in welcher Weise eine bestimmte institutionelle Ausgestaltung des Wissenschaftsbetriebs die heute üblichen Evaluationen zurückdrängen und teilweise sogar ersetzen könnte:

Wenn Universitäten einem stärkeren Wettbewerb unterworfen werden, ist keine *staatliche* Evaluation mehr nötig. Die Studierenden zahlen dann kostendeckende Studiengebühren und suchen sich die Universität aus, die ihrer Ansicht nach die besten Leistungen bietet. Umgekehrt haben die Universitäten die Freiheit, sich die Studierenden auszuwählen, die ihre Anforderungen am besten erfüllen und ihre Reputation verbessern. Der Wettbewerb zwischen den verschiedenen Hochschulen steigert die

Qualität von Ausbildung und Forschung. Beide Seiten müssen sich anstrengen, um ihre Ziele zu erreichen: Wer eine akademische Ausbildung anstrebt, muss sich bemühen, von einer ihm oder ihr zusagenden Hochschule angenommen zu werden. Um bei der Bewerbung erfolgreich zu sein, sind gute Schulnoten vorzuweisen. Das universitäre Auswahlverfahren führt somit dazu, dass sich die jungen Menschen schon auf der Gymnasialstufe wesentlich mehr anstrengen, als dies heute gemeinhin der Fall ist. Bei der Wahl der geeigneten Universität helfen die von privaten Evaluationsagenturen bereitgestellten Informationen über Studienbedingungen (z. B. in Bezug auf die Ausstattung der Bibliotheken oder die Qualität der wissenschaftlichen Betreuung), die Berufsaussichten der Abgänger (darunter auch die zu erwartenden Löhne) und das Prestige einer bestimmten Universität. Für Letzteres ist die Zahl der dort lehrenden hervorragenden Gelehrten wie etwa Nobelpreisträger oder allgemein bekannte Intellektuelle, aber auch die Tradition ausschlaggebend. Wer die wissenschaftliche Laufbahn anstrebt und sich deshalb für eine Graduiertenausbildung interessiert, wird eher Evaluationsberichte zu Rate ziehen, die Aufschluss geben über die Bedingungen eines Promotionsstudiums und die Qualität der Forschung. Die allgemeine Öffentlichkeit fragt wiederum Evaluationsergebnisse nach, die den internationalen Vergleich von Bildungsanstalten ermöglichen. Auch die Universitäten werden privat angebotene Evaluationsergebnisse nachfragen, wobei für sie die Qualität der Ausbildung an den verschiedenen Gymnasien im Vordergrund stehen dürfte. Dies hilft ihnen, besonders begabte und bildungsfähige Absolventen auszuwählen und sich dadurch besonders erfolgreiche Abgänger zu sichern, was sich wiederum positiv auf das Prestige einer Universität auswirkt.

In dem geschilderten System, in dem sowohl den Nachfragenden als auch den Anbietenden unterschiedliche Möglichkeiten zur Auswahl stehen, würde also eine Vielzahl von Ranglisten und Evaluationen entstehen. Auch das Angebot unterläge einem Wettbewerb, der die Anbieter zu sorgfältiger Arbeit zwänge. Denn die genannten Nachfragenden würden keine Evaluationen konsultieren, die ein verzerrtes Bild gäben. Hier gilt ganz Ähnliches wie für Restaurantführer. Niemand würde diese konsultieren, wenn bekannt wäre, dass sich Wirte gegen eine entsprechende Bestechung aufnehmen und hervorragend benoten lassen könnten.

Das skizzierte System des Qualitätswettbewerbs bei den Evaluationen unterscheidet sich grundlegend von einer staatlich verordneten Evaluation, deren Ziel darin besteht, Aufschlüsse über die beste Zuwendung der Budgetmittel zu erhalten. In einem Wettbewerbssystem entsteht die Qualität der Lehre und Forschung als Ergebnis eines Prozesses von unten. Es wird nicht (wie gegenwärtig in Deutschland) der Versuch unternommen, die Elitehochschulen von oben politisch und bürokratisch zu bestimmen. Ebenso wenig wird die eine, allein „richtige" Evaluation eines Faches angestrebt (wie dies zurzeit vom Deutschen Wissenschaftsrat für die Soziologie und Chemie versucht wird). Vielmehr lässt der Wettbewerb eine Vielfalt von Hochschulen entstehen, die sich unterschiedlichen Zwecken widmen, und entsprechend vielfältige Ranglisten. Wie die Verhältnisse in Nordamerika zeigen, führt dieses System keines-

wegs zum Niedergang der Universitäten. Bekanntlich werden gerade die führenden amerikanischen Universitäten wie Harvard, Yale, Princeton oder Stanford privat geführt (vgl. R. Frey 1997).

Dieses Beispiel bezweckt, *Alternativen* zu den heute gängigen staatlich veranstalteten Evaluationen aufzuzeigen, nicht, ein privates Wettbewerbssystem für Universitäten zu befürworten. Es will vielmehr Möglichkeiten darlegen, wie eine Umgestaltung des Universitätssystems die Ziele erreichen kann, die gegenwärtig mit einem unzulänglichen Evaluationssystem angestrebt werden.

5 *Die Alternative der sorgfältigen Personalauswahl*

Die heute üblich gewordenen nachträglichen Evaluationen ganzer Universitäten, Fakultäten, Fachbereiche, Institute und Forschungsteams ließen sich zu einem guten Teil umgehen, wenn die Forschenden und Lehrenden sorgfältig ausgewählt würden. Die Strategie der sorgfältigen Personalauswahl setzt die Ressourcen *zukunftsorientiert* ein, indem Personen bestmöglich mit den zu bewältigenden Aufgaben betraut werden. Das Gewicht wird auf den Auswahlprozess statt auf Überprüfung gelegt.[9] Sobald eine Person einmal ernannt ist – zum Beispiel eine Professur für ein bestimmtes Wissensgebiet erhalten hat – wird ihr vertraut und davon ausgegangen, dass sie die erwarteten Leistungen auch erbringen wird. Man lässt sie ungestört arbeiten. Dabei ist mit einer erheblichen Varianz zu rechnen. Einige unter den ausgewählten Personen werden daraufhin in ihrer Leistung nachlassen und sich nicht mehr stark engagieren, andere hingegen werden durch den gewährten Freiraum beflügelt und erreichen Spitzenleistungen. In der Wissenschaft sollten Letztere zählen und die Unwilligen und Versager als notwendiges Übel betrachtet werden, damit die anderen große und insbesondere innovative Ergebnisse erzielen können. James Bryan Conant, der bedeutende Präsident der Harvard Universität, zählte zu den entschiedenen Fürsprechern für eine derartige Organisation der Wissenschaft: "There is only one proved method of assisting the advancement of pure science – that is picking men of genius, backing them heavily, and leaving them to direct themselves." (Letter to the New York Times, 13. August 1945, zit. in Renn 2002: 28).

Die gleiche Auffassung findet sich auch noch heute in den „Principles Governing Research at Harvard" (http://www.fas.harvard.edu/research/greybook/principles.html), wo festgehalten wird: "The primary means for controlling the quality of scholarly activities of this Faculty is through the rigorous academic standards applied in selection of its members."

Im Gegensatz dazu vermag die ständige Evaluation der Leistungen von Forschenden im Grunde nichts anderes, als ein bestimmtes Durchschnittsniveau zu sichern; die als Kontrolle erlebten fortwährenden Beurteilungen führen entsprechend zu „normaler" Wissenschaft ohne Spitzenleistungen. Diese Situation wird noch da-

9 Zu einem analogen Auswahlprozess in der Politik vgl. Besley (2005), Cooter (2002).

durch verstärkt, dass es – wie oben ausgeführt – so gut wie unmöglich ist, sich anstehenden Evaluationen zu entziehen. Schwer vorstellbar, dass in der heutigen akademischen Welt mit ihren dauernden Evaluationen Ausnahmeforscher in den Naturwissenschaften wie Einstein und Planck oder Keynes und Hicks in der Wirtschaftsforschung hätten prosperieren können. Nicht genug damit, dass sie durch die notwendigen Rechtfertigungen ihrer Tätigkeit („Was haben Sie im letzten Halbjahr geforscht und veröffentlicht?") von ihrer Forschungstätigkeit abgehalten worden wären, sie hätten überdies zu ihrer Zeit in einer Evaluation womöglich sogar schlecht abgeschnitten, weil sie die Prinzipien und Normen der „normalen" Wissenschaft (Kuhn 1962) in Frage stellten und verwarfen. Für sehr viele bahnbrechende Erkenntnisse gilt, dass sie von den Zeitgenossen nicht verstanden und als lächerlich bezeichnet werden. Ein Beispiel dafür ist Freges innovative mathematische Theorie aus dem Jahr 1897, die bei der Veröffentlichung in fünf von sechs Besprechungen extrem herablassend beurteilt wurde. Es dauerte zwanzig Jahre, bis sie allmählich (u. a. von Bertrand Russell) in ihrer Bedeutung erfasst wurde, und erst in den 1950er Jahren erhielt sie die ihr gebührende Anerkennung. Auch Semmelweis' Erkenntnisse zur Antisepsis (1847) wurden erst nach rund zwanzig Jahren akzeptiert, und die fundamentalen astrologischen Einsichten von Kopernikus (1473-1543) sogar noch fünfzig bis sechzig Jahre nach seinem Tod von anderen Astronomen für absurd gehalten (siehe ausführlich Gillies 2005).

Zu einer nicht selten als Bevormundung empfundenen Evaluation gibt es also die Alternative einer sorgfältigen Personalauswahl und des Vertrauens in den Willen und die Fähigkeit zur Leistung der Einzelnen.[10] Eine Bewertung der Leistung der Forschenden vollzieht sich dann gleichsam von alleine, in einem dezentralen, autonomen und zuweilen langsamen Wissenschaftsprozess. Es ist daran zu erinnern, dass es gerade dieses System ist, dem die deutschsprachige Wissenschaft in der Vergangenheit ihre Weltgeltung zu verdanken hatte. Soll es heute durch das System unablässiger Evaluationen ersetzt werden, dann müssen überzeugende Argumente vorgebracht werden, warum es seine Wirksamkeit eingebüßt haben soll.

6 *Abschließende Bemerkungen*

Evaluationen im Sinne einer nachträglichen Bewertung der Leistung von Institutionen und Personen durch außenstehende Gutachtende vor allem zum Zwecke der Mittelzuweisung weisen einige „verborgene" Kosten auf. Dazu zählen vor allem schädliche Anreizverzerrungen, eine induzierte Verkrustung und ein verfehlter Entscheidungsansatz. Weil diese Kosten gewöhnlich unberücksichtigt bleiben, gelten Evaluationen als „Allheilmittel" und werden zu oft und zu intensiv angewandt. Im vorliegenden Beitrag wird nicht gegen Evaluationen an sich argumentiert, wohl aber gegen ihre Dominanz

10 Welche Auswirkungen Vertrauen im Unterschied zu Kontrolle hat, wird analysiert bei Bohnet/Frey/Huck (2001) und Huang (2005).

und Allgegenwärtigkeit. Es handelt sich bei der vorliegenden kritischen Auseinandersetzung mit Evaluationen auch nicht um ein Plädoyer für deren (methodische) Optimierung. Die hier vorgebrachten Einwände sind vielmehr grundsätzlicher Art und können auch durch differenziertere Evaluationen nicht einfach beseitigt werden. Ganz im Gegenteil: Es ist sogar denkbar, dass verbesserte intensivere Evaluationen die hier aufgeführten fundamentalen Probleme nur noch verschlimmern.

Die häufig vorgebrachte Ansicht, es gäbe keine Alternativen zu Evaluationen, wird verworfen. Stattdessen wird die Möglichkeit institutioneller Änderungen und sorgfältiger Personalauswahl hervorgehoben. Die Debatte sollte sich nicht ausschließlich mit den Vorzügen und Grenzen von Evaluationen befassen, sondern auch ernsthaft andere Möglichkeiten der Gewährleistung von Exzellenz einbeziehen.

Literatur

Amabile, Teresa (1996): Creativity in Context: Update to the Social Psychology of Creativity. Boulder: Westview Press

Amabile, Teresa (1998): How to kill creativity. In: Harvard Business Review 76(5): 76-87.

Backes-Gellner, Uschi/Petra Moog (Hg.) (2004): Ökonomie der Evaluation von Schulen und Hochschulen. Berlin: Duncker und Humblot.

Bedeian, Arthur G. (2003): The manuscript review process: The proper roles of authors, referees, and editors. In: Journal of Management Inquiry 12: 331-338.

Bénabou, Roland/Jean Tirole (2003): Intrinsic and extrinsic motivation. In: Review of Economic Studies 70(3): 489-520.

Besley, Timothy (2005): Political selection. In: Journal of Economic Perspectives 19: 43-60.

Bohnet, Iris/Bruno S. Frey/Steffen Huck (2001): More order with less law: On contract enforcement, trust, and crowding. In: American Political Science Review 95(1): 131-144.

Bräuninger, Michael/Justus Haukap (2003): Reputation and relevance of economics journals. In: Kyklos 56: 175-198.

Broadfoot, Patricia M. (1996): Education, Assessment and Society. Buckingham: Open University Press.

Brook, Richard (2002): The Role of Evaluation as a Tool for Innovation in Research. In: Max Planck Forum 5, Innovative Structures in Basic Decision Research. Ringberg Symposium, 4.-7. Oktober 2000 in München: 173-179.

Brück, Tilman/Andreas Stephan (2006): Do Eurozone countries cheat with their budget deficit forecasts? In: Kyklos 59: 3-16.

Cameron, Judy/Katherine M. Banko/W. David Pierce (2001): Pervasive negative effects of rewards on intrinsic motivation: The myth continues. In: The Behavior Analyst 24: 1-44.

Cash, David/William C. Clark (2001): From Science to Policy: Assessing the Assessment Process. KSF Faculty Research Working Papers Series RWP01-045.

Chapman, Chris (2006). Joining accountabilty and autonomy in research. In: Foresight Europe 2 (March): 13-14.

Chrystal K. Alec/Paul D. Mizen (2003): Goodhart's Law: Its origins, meaning and implications for monetary policy. In: Paul D. Mizen (Hg.): Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart. Vol. 1. Cheltenham, U.K./Northampton, MA, USA: Edward Elgar: 221-243.

Cooter, Robert D. (2002): Who Gets on Top in Democracy? Elections as Filters. Working Paper Series No. 74. Berkeley Online Program in Law and Economics.

Daily, Catherine M./Dan R. Dalton/Albert. A. Cannella (2003): Introduction to special topic forum. Corporate governance: Decades of dialogue and data. In: Academy of Management Review 28(3): 371-382.

Daniel, Hans-Dieter (1993): Die Wächter der Wissenschaft. Weinheim: Wiley-VCH.

Daniel, Hans-Dieter/Rudolf Fisch (Hg.) (1988): Evaluation von Forschung: Methoden, Ergebnisse, Stellungnahmen. Konstanz: Universitätsverlag.

Day, Patricia/Rudolf Klein (1990): Age of Inspection. Inspecting the Inspectors. London: Rowntree Foundation.

De Bruijn, Hans (2002): Managing Performance in the Public Sector. London/New York: Routledge.

Deci, Edward L./Richard Koestner/Richard M. Ryan (1999): A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. In: Psychological Bulletin 125(6): 627-668.

Economist (2002): Ranking Business Schools. The Numbers Game. 12. Oktober: 69.

Erlei, Mathias/Martin Leschke/Dirk Sauerland (1999): Neue Institutionenökonomik. Stuttgart: Schäffer-Poeschel.

European Institute for Advanced Studies in Management (Hg.) (2006): Accountability in research. In: Foresight Europe 2.

Fehr, Ernst/Simon Gächter (2002): Do Incentive Contracts Crowd Out Voluntary Cooperation? Institute for Empirical Research in Economics, Working Paper No. 34.

Figlio, David/Lawrence Getzler (2003): Accountability, Ability and Disability: Gaming the System. NBER Working Paper No 9307.

Fischer, Klaus (1998): Evaluation der Evaluation. In: Wissenschaftsmanagement 5: 16-21.

Frey, Bruno S. (1990): Ökonomie ist Sozialwissenschaft: Die Anwendung der Ökonomie auf neue Gebiete. München: Vahlen.

Frey, Bruno S. (1992): Tertium datur: Pricing, regulation and intrinsic motivation. In: Kyklos 45: 161-184.

Frey, Bruno S. (1997): Not Just for the Money: An Economic Theory of Personal Motivation. Cheltenham, U.K.: Edward Elgar.

Frey, Bruno S. (2001): Inspiring Economics: Human Motivation in Political Economy. Cheltenham, U.K.: Edward Elgar.

Frey, Bruno S. (2003). Publishing as prostitution? – Choosing between one's own ideas and academic success. In: Public Choice 116: 205-223

Frey, Bruno S./Reto Jegen (2001): Motivation crowding theory. In: Journal of Economic Surveys 15(5): 589-611.

Frey, Bruno S./Gebhard Kirchgässner (2002): Demokratische Wirtschaftspolitik. 3. Aufl., München: Vahlen.

Frey, Bruno S./Margit Osterloh (2000a): Pay for performance – Immer empfehlenswert? In: Zeitschrift für Führung und Organisation (ZFO) 69: 64-69.

Frey, Bruno S./Margit Osterloh (Hg.) (2000b): Managing Motivation: Wie Sie die neue Motivationsforschung für Ihr Unternehmen nutzen können. Wiesbaden: Gabler.

Frey, Bruno S./Margit Osterloh (2005): Yes, managers should be paid like bureaucrats. In: Journal of Management Inquiry 14: 96-111.

Frey, René L. (1997): Universitäten im Aufbruch. Volkswirtschaftliche Analyse der gegenwärtigen Reformen. Rektoratsrede gehalten an der Jahresfeier der Universität Basel, Basler Universtitätsreden 93. Basel: Helbing und Lichtenhahn.

Gibbons, Robert (1998): Incentives in organizations. In: Journal of Economic Perspectives 12: 115-132.

Gillies, Donald (2005): Lessons from the History and Philosophy of Science Regarding the Research Assessment Exercise. Paper read at the Royal Institute of Philiosophy on 18 November 2005. (www.ucl.ac.uk/sts/gillies).

Gillies, Donald (2006): Why research assessment exercises are a bad thing. In: Post-Autistic Economics Review 37: 2-9.

Heckhausen, Heinz (1989): Motivation und Handeln. 2. Aufl., Berlin etc.: Springer.

Holcombe, Randall G. (2004): The national research council ranking of research universities: Its impact on research in economics. In: Econ Journal Watch 1(3): 498-514.

Holmstrom, Bengt/Paul Milgrom (1991): Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. In: Journal of Law, Economics, and Organization 7(2): 24-52.

Hopwood, Anthony G. (2005): Editorial: After 30 years. In: Accounting, Organization and Society 30: 585-586.

Huang, Fali (2005): To Trust or to Monitor: A Dynamic Analysis. Mimeo, School of Economics and Social Sciences: Singapore Management University.

Jordan, Thomas Edward (1989): Measurement and Evaluation in Higher Education: Issues and Illustrations. London: Falmer Press.

Kirchgässner, Gebhard (2000): Homo oeconomicus. 2. Aufl., Tübingen: Siebeck.

Klostermeier, Johannes (1994): Hochschul-Ranking auf dem Prüfstand: Ziele, Methoden und Möglichkeiten. Interdisziplinäres Zentrum für Hochschuldidaktik der Universität Hamburg.

Knorr Cetina, Karin (2006): Knowledge cultures. In: Foresight Europe 2(March): 7-11.

Kogut, Bruce (2006): Accountability in research: An introduction to the issue (and issues). In: Foresight Europe 2(March): 3-5.

Kozar, Gerhard (1999): Hochschul-Evaluierung: Aspekte der Qualitätssicherung im tertiären Bildungsbereich. Wien: WUF.

Kuhn, Thomas S. (1962): The Structure of Scientific Revolution. Chicago: University of Chicago Press.

Lindsay, Douglas (1989: Using citation counts as a measure of quality in science measuring What's measurable rather than what's valid. In: Scientometrics 15: 189-203.

Max-Planck-Gesellschaft (2002): Innovative Structures in Basic Decision Research. Ringberg Symposium, 4.-7. Oktober 2000 in München.

McCabe, Donald L./Linda Klebe Trevino/Kenneth D. Butterfield (1996): Cheating in academic institutions: A decade of research. In: Ethics and Behavior 11: 219-232.

Neave, Guy (1988): On the cultivation of quality, efficiency and enterprise: An overview of recent trends in higher education in Western Europe, 1986-1988. In: European Journal of Education 23(1-2): 7-23.

Osterloh, Margit/Bruno S. Frey (2005): Shareholders Should Welcome Employees as Directors. IEW Working Paper No. 228, Institute for Empirical Research in Economics: University of Zurich.

Power, Michael (1994): The Audit Explosion. London: Demos.

Power, Michael (1997): The Audit Society. Ritual of Verification. Oxford: Oxford University Press.

Renn, Jürgen (2002): Challenges from the past. Innovative structures for science and the contribution of the history of science. In: Max Planck Forum 5, Innovative Structures in Basic Decision Research. Ringberg Symposium, 4.-7. Oktober 2000 in München: 25-36.

Richter, Rudolf/Erik Furubotn (1999): Neue Institutionenökonomik. Tübingen: Siebeck.

Ridgway, V.F. (1956): Dysfunctional consequences of performance measurement. In: Administrative Science Quarterly 1: 240-247.

Röbbecke, Martina/Dagmar Simon (1999): Zwischen Reputation und Markt – Ziele, Verfahren und Instrumente von (Selbst)Evaluationen außeruniversitärer, öffentlicher Forschungseinrichtungen. WZB-Discussion Paper: 99-601.

Röbbecke, Martina/Dagmar Simon (2001): Assessment of the Evaluation of Leibniz-Institutes – External Evaluation and Self-Evaluation. In: Philip Shapira/Stefan Kuhlmann (Hg.): Proceeding from the 2000 US-EU Workshops on Learning from Science and Technology Policy Evaluation. Bad Herrenalb, Kap. 8: 16-23.

Royal Netherlands Academy of Arts and Sciences (2005): Judging Research on its Merits. Amsterdam.

Russon, Craig/Karen Russon (Hg.) (2000): The Annotated Bibliography of International Programme Evaluation. Dordrecht: Kluwer.

Schulze, Günther/Björn Frank (2003): Deterrence versus intrinsic motivation: Experimental evidence on the determinants of corruptibility. In: Economics of Governance 4: 143-160.

Starbuck, William H. (2004): Methodological challenges posed by measures of performance. In: Journal of Management and Governance 8: 337-343.

Stockmann, Reinhard (Hg.) (2004): Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder. 2. Aufl., Opladen: Leske + Budrich.

Suvorov, Anton/Jeroen van de Ven (2006): Discretionary Rewards as a Feedback Mechanism. (Available at SSRN: http://ssrn.com/abstract=889280).

Tucci, Christopher (2006): Why Europe will never have accountability in research. In: Foresight Europe 2 (March): 27-29.

Weingart, Peter (2005): Impact of bibliometrics upon the science system: Inadvertent consequences? In: Scientometrics 62: 117-131.