Perspektiven der Wirtschaftspolitik 2007 8(3): 207–220

Evaluierungen, Evaluierungen ... Evaluitis

*Bruno S. Frey**
Universität Zürich und CREMA – Center for Research in Economics, Management and the Arts

1. Evaluitis

In der Wissenschaft wird immer häufiger und immer umfassender evaluiert. In immer kürzeren Abständen werden ganze Universitäten und Disziplinen, Fakultäten, Fachbereiche, Institute, Forschungsgruppen und einzelne Forschende begutachtet. Das Fortschreiten von Evaluationen scheint unaufhaltsam und kann mit einer Modewelle verglichen werden, bei der sich ohne viel Überlegung alle beteiligen, die nicht als rückständig gelten wollen. Etwas weniger freundlich kann von einer sich epidemisch ausbreitenden Krankheit gesprochen werden – der „Evaluitis" – die ganz besonders auch die Wissenschaft befallen hat. Evaluationen und daraus abgeleitete Rankings sind heute in der Tat zu einem Allgemeingut geworden. Entsprechend wird von einer „Audit Explosion" (Power, 1994), einer „Audit Society" mit ihren "Rituals of Verification" (Power, 1997), von der „Age of Inspection" (Day und Klein, 1990) oder vom „Evaluative State" (Neave, 1988) gesprochen.

Selbstverständlich wird hier nicht gegen die Abwägung von Nutzen und Kosten im Allgemeinen argumentiert. Sie gehört zum Kern ökonomischen Denkens und ist unverzichtbar. Dieser Beitrag möchte hingegen auf einige wenig diskutierte, *verborgene und damit gewöhnlich vernachlässigte Kosten von Evaluationen* aufmerksam machen. Unter „Evaluation" wird hier eine *nachträgliche Einschätzung der Leistung einer Organisation oder Person durch von außen kommende Experten verstanden*. Entscheidend dabei ist die Einschränkung auf *nachträgliche* Einschätzung und von *außen kommend*. Wenn also eine Universität oder Firma die Fähigkeiten von Bewerbenden für eine Stelle einschätzt und dabei die zukünftige Kapazität abzuschätzen versucht, handelt es sich *nicht* um eine

*Institut für Empirische Wirtschaftsforschung der Universität Zürich, Winterthurerstrasse 30, CH-8006 Zürich, Schweiz. E-Mail: bsfrey@iew.unizh.ch. – Ich bedanke mich für wertvolle Hinweise bei Michael Baurmann, Reiner Eichenberger, Simon Lüchinger und Margit Osterloh, sowie beim Gutachtenden für seine/ihre ausführlichen und hilfreichen Bemerkungen. Einige Einsichten verdanke ich meiner eigenen Tätigkeit als Evaluator verschiedener Forschungseinrichtungen in unterschiedlichen Ländern.

Evaluation im hier definierten Sinne. Vielmehr geht es in dieser Arbeit um formale Evaluationen, bei denen die vergangene Leistung einer Einrichtung oder Person von außen stehenden Fachleuten eingeschätzt wird.

Die hier verwendete Definition entspricht sowohl dem Alltags- als auch dem Wissenschaftsverständnis; so sagt etwa Brook (2002, S. 173): „By evaluation, I shall mean the situtation where visiting experts come from outside your organization or system and say what they think about it". Diese Arbeit konzentriert sich auf Evaluationen in der Wissenschaft im staatlichen Auftrag, die insbesondere helfen sollen, die geeignete Zuteilung finanzieller Mittel zu unterstützen.

In soweit Kosten nicht berücksichtigt werden, wenn darüber entschieden wird, ob eine Evaluation durchgeführt werden sollte (falls darüber überhaupt noch explizit entschieden wird), wird der Nettonutzen dieses Instrumentes *systematisch* überschätzt. Das sich einstellende Gleichgewicht, in dem sich Grznutzen und die Grenzkosten der Evaluation entsprechen, ist nicht optimal, weil gewichtige Grenzkosten vernachlässigt werden. In diesem Falle werden *zu viele* und *zu intensive* Evaluationen durchgeführt als gesellschaftlich sinnvoll wäre. Insofern lässt sich „Evaluitis" als eine Krankheit bezeichnen (Frey und Osterloh, 2006).

Zur Vermeidung von Missverständnissen sei noch einmal betont, dass es in diesem Aufsatz vor allem darum geht, vernachlässigte Kosten der Evaluation zu identifizieren, weniger jedoch darum, inwiefern der Markt funktioniert und damit Evaluationen unnötig sind. Ebenso wenig geht es um die Frage, wer diese Evaluationen durchführen sollte. Diskutiert wird vielmehr, welche Kosten typischerweise gering geschätzt oder völlig vernachlässigt werden. Damit wird *kein* Argument gegen Evaluationen *an sich* vorgebracht; in manchen Fällen sind sie durchaus sinnvoll. Allerdings wird nicht die Auffassung geteilt, die heutigen Evaluationen seien zwar mangelhaft und sollten deshalb einfach verbessert werden. Die hier vorgebrachten Einwände sind grundsätzlich und lassen sich nicht einfach beseitigen, indem die Evaluationen differenzierter und genauer werden. Verbesserte Evaluationen senken die berücksichtigten Grenzkosten der Anwendung dieses Instruments und führen zu einer noch intensiveren Anwendung. Damit können sich die hier aufgeführten fundamentalen Probleme sogar noch verschlimmern.

Die riesige Literatur zum Thema Evaluation wird hier nur beispielhaft aufgeführt.[1] In diesem Beitrag wird somit nicht auf die sattsam bekannten Kosten auf Seiten der Evaluierer und der Evaluierten eingegangen, insbesondere auf den erheblichen zeitlichen und materiellen Aufwand. Im Zentrum stehen

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

somit gerade bei der Anwendung in der Praxis vernachlässigte Aspekte. Es wird nicht behauptet, dass diese nirgendwo erwähnt würden, wohl aber dass sie kaum oder gar nicht beachtet werden. 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.[2] Auf Evaluationen kann nicht vollständig verzichtet werden, jedoch können sie auf ein geringes Maß reduziert werden, wenn *mittels geeigneter Institutionen geeignete Anreize* vermittelt werden und wenn das Schwergewicht auf eine sorgfältige *vorherige* Auswahl von Personen gelegt wird.

Abschnitt 2 befasst sich mit der durch Evaluationen verursachten Anreizverzerrung bei den Evaluierten, Abschnitt 3 mit der induzierten Verkrustung und Abschnitt 4 mit dem verfehlten Entscheidungsansatz und damit dem geringen Nutzen für die Entscheidungsbildung. Im zweiten Teil werden die Alternativen zu den Evaluationen behandelt. Abschnitte 5 und 6 argumentieren, dass ein gewünschtes Verhalten auch mittels institutionellen Änderungen und einer sorgfältigen Personenauslese erzielt werden kann. Im letzten Abschnitt werden Folgerungen gezogen.

2. Evaluation verzerrt Anreize

Das Instrument der Evaluation verändert das Verhalten der davon betroffenen Personen in systematischer und unbeabsichtigter Weise: (a) Alle Beteiligten konzentrieren sich allein auf das, was *gemessen* wird; (b) die *intrinsischen Arbeitsanreize werden verdrängt*; und (c) die verwendeten Kennziffern werden *manipuliert*.

2.1 Nur was gemessen wird, zählt.

Das Phänomen des „Multi-tasking" ist seit einigen Jahren intensiv diskutiert worden (z.B. Holmstrom und Milgrom, 1991, Gibbons 1998, Daily, Dalton und Cenella, 2003, Suvorov und van de Ven, 2006). Die Vorgesetzten (Prinzipale) legen die Maßstäbe fest, mit denen die Leistung einer Institution oder einer Person beurteilt wird. Für keine Tätigkeit – außer möglicherweise einfachster Fließbandtätigkeit – lassen sich jedoch *alle* relevanten Aspekte definieren und messen. Jede Person hat deshalb die Tendenz, oder wird sogar gezwungen, sich ausschließlich nach den gemessenen Kriterien zu richten und alles andere beiseite zu lassen. In den vielen Fällen, in denen nur Inputs erfasst werden, ist die Verzerrung besonders gewichtig, weil dann die Produktivität völlig vernachlässigt wird.

Das Multi-tasking-Problem hat in der Wissenschaft besonders starke Auswirkungen. So wird heute häufig die Anwerbung von Drittmitteln als

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

Leistungskriterium verwendet (so etwa bei den Organisationen in der Leibniz-Gesellschaft, wozu auch die Max-Planck-Gesellschaften gehören). Damit wird weder der Sinn noch die Produktivität der damit finanzierten Forschung erfasst, vielmehr werden diejenigen Personen belohnt, die besondere Fähigkeiten zum „Verkauf" ihrer Aktivitäten im wissenschaftlichen Raum aufweisen. In der Tat haben in den letzten Jahren mehr oder weniger spezialisierte Wissenschaftsmanager an Bedeutung gewonnen. Jede Person, die mit dem heutigen Wissenschaftsbetrieb vertraut ist, weiß, dass solche Wissenschaftsmanager nicht notwendigerweise auch die besten Wissenschaftler sind. In vielen Fällen existiert vor allem wegen der begrenzten Zeit und auch unterschiedlichen Fähigkeiten ein Trade-off zwischen den beiden Aktivitäten.

Die Anwerbung von Drittmitteln ist vor allem deshalb ein gängiges Maß für erfolgreiche Wissenschaft geworden, weil Geldströme besonders leicht messbar sind. Wenn aber eine wissenschaftliche Einheit damit beurteilt wird, ist sie gezwungen, sich um Drittmittel zu bemühen und dafür schlechter 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 Lindsey (1989, S. 200): „Citation counts as a measure of quality may often be measuring what is measurable rather than what is valid". Vernachlässigt wird dabei die Übertragung wissenschaftlicher Erkenntnisse in die Praxis mittels Publikationen in populären Organen, allgemein bildende Vorträge, Beratungstätigkeit, Betreuung des wissenschaftlichen Nachwuchses, 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. Es wird versucht, die entsprechenden Aspekte ebenfalls quantitativ zu erfassen. Dies wird jedoch *nie* im vollen Umfang möglich sein. Das Multi-tasking-Problem wird deshalb auf immer schwerer messbare Aspekte verlagert, die Verzerrung der Anreize aber nicht beseitigt. Vielmehr kommt es zu einem dauernden Wettlauf zwischen den Evaluierten und den Evaluierern. Das Ergebnis sind immer aufwendigere Evaluationsprozesse, die den Forschenden immer weniger Zeit zur eigentlichen Tätigkeit übrig lassen: „Success in the evaluation process can become a more significant target than success in research itself" (Brook, 2002, S. 176).

Die Erfassung von Zitierungen führt selbst dann zu Verzerrungen im Verhalten, wenn sie vollständig erfasst würden. Sobald die Forschenden wissen, dass ihre wissenschaftliche Leistung nach diesem Kriterium gemessen wird, werden sie sich solchen Forschungsfragen widmen, die gerade Mode sind und deshalb viele Zitierungen versprechen.

2.2 Intrinsische Arbeitsanreize werden verdrängt.

Die mit der Evaluation einhergehende Messung und Beurteilung der Leistung wird von den Betroffenen in der Regel als *kontrollierend* empfunden, was die Arbeitsmotivation beeinträchtigt. Dieser Effekt ist in der Sozialpsychologie

in Hunderten von Laborexperimenten analysiert worden (für umfassende Metastudien siehe Deci, Koestner und Ryan, 1999 und Cameron, Banko und Pierce, 2001). Er ist in der Ökonomik als „Verdrängungseffekt" (Frey, 1992, 1997, Fehr und Gächter, 2002, Bénabou und Tirole, 2003) empirisch anhand von Felduntersuchungen bestätigt worden (eine Übersicht geben Frey und Jegen, 2001). Die Gesamtleistung vermindert sich nicht notwendigerweise. So vermeldet Brook (2002, S. 176) für die britische „Research Assessment Exercise": „. . . we can safely say that the average activity has increased" – zumindest in der evaluationerfassten Dimension. Ob allerdings die Auswirkungen auf die Qualität und Originalität der Forschung günstig waren, muss bezweifelt werden. Wie Amabile (1996, 1998) gezeigt hat, ist die intrinsische Motivation für innovative wissenschaftliche Arbeit von entscheidender Bedeutung. Hinzu kommt, dass gerade Bahn brechende Forschung gegen den herrschenden Konsens in der Wissenschaft verstößt und deshalb in einer Evaluation gerade durch die „Peers" schlecht abschneidet. Historische Untersuchungen (Fischer, 1998, Gillies, 2006) belegen, dass viele besonders wichtige Forschungsergebnisse dem Zeitgeist (im Sinne der „normal science" von Kuhn, 1962) widersprachen und deshalb in einer Evaluation ungünstig beurteilt worden wären.

Eine Evaluation verdrängt die intrinsische Forschungsmotivation nicht notwendigerweise; wird sie von den Betroffenen als *unterstützend* erlebt, wird sie sogar gesteigert (z.B. Heckhausen, 1989). Das gleiche gilt, wenn die Evaluierten die ihnen zukommende Aufmerksamkeit genießen und sich kurzfristig mehr anstrengen (Hawthorne-Effekt). Die beiden Bedingungen dürften zutreffen, wenn die Evaluation neu eingeführt wird. Wird sie jedoch eine unablässige Übung, wird sie als kontrollierend empfunden und die intrinsische Forschungsmotivation wird immer mehr verdrängt.

2.3 Die Leistungskriterien werden manipuliert.

Wenn ein Indikator für die eigene Position wichtig wird, entsteht ein starker Anreiz, diesen Indikator zu seinen eigenen Gunsten zu beeinflussen. Dieser allgemeine Zusammenhang ist mit „Goodhart's Law" (1975) und der „Lucas Critique" (1976) verwandt. Beide sind empirisch auf der Makroebene gut nachgewiesen (vgl. z.B. Chrystal and Mizen, 2003, Brück und Stephan, 2006), gelten aber auch auf der Mikroebene. Schulleitungen können zum Beispiel ihre Beurteilung beeinflussen, indem sie die Schüler auf die Examensaufgaben hin trainieren („teaching-to-the-test"), schlechte Schüler unter irgendwelchen Vorwänden von den entsprechenden Tests ausschließen und damit die Ergebnisse ihrer Schule künstlich verbessern (für empirische Evidenz für die Vereinigten Staaten vgl. Figlio und Getzler, 2003). Manager beeinflussen die Leistungskriterien, sobald ihr Einkommen davon abhängig ist. So treiben sie (kurzfristig) die Aktienpreise in die Höhe, wenn ein Teil ihres Gehaltes in der Form von Aktienoptionen ausgerichtet wird (z.B. Osterloh und Frey, 2005, Frey und Osterloh, 2000, 2005).

Eine derartige Manipulation hat sich auch in der Wissenschaft verbreitet, seit die Forschungsleistung im Zuge von Evaluationen anhand der Zahl der Publikationen und Zitierungen gemessen wird. Für die Wissenschaftskultur noch schädlicher ist das Hochjubeln von Ergebnissen in der Forschung weit über deren Bedeutung hinaus. So besteht ein verstärkter Anreiz, nur noch erfolgreiche Tests zu publizieren und die negativen Ergebnisse zu verschweigen oder sogar zu beseitigen. Noch weiter gehender ist der Anreiz zum Betrug mittels Fälschung von Forschungsergebnissen. In einem Experiment wurde gezeigt, dass sich kontrolliert fühlende Personen in weit stärkerem Ausmaß bereit sind, zu betrügen (Schulze und Frank, 2003). Dass dies auch tatsächlich im Wissenschaftsbetrieb vorkommt, zeigen verschiedene Skandale der letzten Zeit (vgl. z.B. McCabe, Trevino und Butterfield, 1996, Bedeian, 2003, Frey 2003).

Selbstverständlich lassen sich nicht alle Leistungsindikatoren in gleichem Ausmaß manipulieren. Allerdings wäre es verfehlt, davon auszugehen, dass sich bestimmte Indikatoren grundsätzlich nicht verfälschen ließen. Es muss immer berücksichtigt werden, dass wenn Indikatoren wichtig werden (insbesondere auch für das Gehalt und die zugewiesenen Mittel), starke Anreize zur Manipulation entstehen. In der Wissenschaft tätige Personen sind nicht weniger als andere Leute fähig, Indikatoren auf innovative Weise zu verfälschen.

3. Evaluationen werden zu einem Selbstzweck

Evaluationen bewirken eine *Verkrustung* sowohl (a) auf Seiten der Evaluierten wie auch (b) der Evaluierer. Wenn sich die Bedingungen ändern, insbesondere wenn es sich herausstellt, dass Evaluationen weniger erfolgreich sind als bisher angenommen, verhindern starke Kräfte, dass die Zahl, Häufigkeit und Intensität der Evaluationen vermindert wird. Diese Beharrungstendenzen sind wegen der inzwischen weit entwickelten Institutionalisierung von Evaluationen besonders ausgeprägt.

3.1 Die Situation der Evaluierten

Die Angehörigen einer Institution, oder einzelne Forschende, für die eine Evaluation vorgesehen ist, können sich nicht gegen deren Durchführung wenden. Dies gilt selbst wenn sie überzeugt sind, 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. Es würde ihnen sofort vorgeworfen, sie hätten Angst vor dem Ergebnis der Evaluation. Da die Evaluation typischerweise mit einer Mittelvergabe einhergeht, müssen sie sich wider bessere Einsicht an der Evaluation beteiligen. Sie tun deshalb gut daran, begeistert mitzumachen. Damit wird der Anschein erweckt, die Evaluierten seien von den Vorzügen einer Evaluation überzeugt und mit dem Einsatz von Evaluationen einverstanden. Einer zynischen Haltung zur Wissenschaft und deren Ergebnissen wird damit Vorschub geleistet.

3.2 Die Situation der Evaluierer

Die Institutionen und Personen, welche die Evaluation durchführen, haben ein direktes Einkommens- und Karriereinteresse. Besonders ausgeprägt ist dieses Interesse bei privaten Anbietern, aber auch staatlichen Institutionen, deren Bedeutung und deshalb auch Budgetzuweisungen vom Fortbestand des Evaluierens abhängt. Sie sind deshalb bestrebt, Evaluationen auf immer weitere Bereiche auszudehnen, zu intensivieren und in immer kürzeren Abständen durchzuführen. Die negativen Aspekte von Evaluationen werden heruntergespielt. Gleichzeitig wird den im zweiten Teil dieses Aufsatzes genannten Alternativen zur Evaluation wenig oder gar keinen Raum gegeben.

4. Der Nutzen von Evaluationen für Entscheidungen ist gering

In der Regel wird als selbstverständlich unterstellt, dass die durch die Evaluation gewonnene Information wesentlich dazu beiträgt, die Entscheidungen über die wissenschaftliche Forschung zu verbessern. Dies gilt jedoch weniger als gemeinhin erwartet, denn (a) der Informationsgewinn ist bescheiden und es werden (b) weitgehend irrelevante Informationen gesammelt.

4.1 Wenig Informationsgewinn

In der Wissenschaft ist in der „Scientific Community" sehr wohl bekannt, welche Institutionen und Personen gute Forschung betreiben. Viele dieser Informationen werden heute rasch durch Wissenschaftsjournalisten der Öffentlichkeit vermittelt und sind den über die Ressourcenzuweisung entscheidenden Beamten und Politikern geläufig. Sie brauchen deshalb nicht unbedingt auf eine formelle Evaluation zu warten, um vernünftig entscheiden zu können. Evaluationen dienen in der Tat häufig nur dazu, bestehendes Wissen zu bestätigen, dienen also vor allem bürokratischen Zwecken.

Bestätigen Evaluationen das bereits geläufige Wissen, ist somit wenig oder nichts gewonnen. Kommt hingegen eine Evaluation zu einer ungünstigen Einschätzung einer als gut geltenden Forschung, wird das Ergebnis zu Recht angezweifelt. Das gleiche gilt, wenn die Evaluation zu einem guten Ergebnis kommt, wenn in der „Gelehrtenrepublik" das Gegenteil fest steht. 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 entsprechend höhere Budgetzuweisungen. Wer hingegen schlecht eingeschätzt wird, wird große Anstrengungen unternehmen, sich gegen die Auswirkungen zu wehren. Negative Evaluationsergebnisse haben selten große Auswirkungen. Oft sind sie nur symbolischer Natur, ein Effekt, der anderswie weit

kostengünstiger erreicht werden könnte. Nur wenn in der Scientific Community keine Einigkeit über die Qualität eines Forschenden oder einer Forschungsorganisation besteht, verbessert eine Evaluation die Information. Aber gerade dann sind die Ergebnisse besonders umstritten und haben kaum Bestand.

4.2 Die gewonnenen Informationen sind für die Entscheidungen wenig relevant.

Evaluationen suchen in aller Regel das bestehende *Niveau* einer Leistung anhand von Indikatoren wie etwa Publikations- und Zitierhäufigkeit oder Lehrerfolg zu erfassen. Für politische Entscheidungen sind jedoch diese Informationen von wenig Bedeutung, denn es bleibt völlig offen, was daraus zu schließen ist. Sollte den für schlecht befundenen Institutionen und Forschenden die Mittel gekürzt werden? Oder sollte ihnen nicht *zusätzliche* Mittel bewilligt werden, damit sie ihre Qualität erhöhen können? Sollte umgekehrt den für gut befundenen Institutionen und Forschenden die Mittel gekürzt werden, weil sie ja ohnehin erfolgreich sind? Diese Fragen lassen klar erkennen, dass in der politischen Auseinandersetzung um Mittel noch alles völlig offen ist. Auch aus normativ-ökonomischer Sicht ist es offen, ob zum Beispiel eine Wissenschaftseinheit, die in der Vergangenheit besonders erfolgreich war, auf Grund einer Mittelerhöhung in der Zukunft *deshalb* mehr leisten wird als eine Institution, die bisher nicht besonders erfolgreich war. Das „Gesetz" des abnehmenden Grenznutzens wird auch bei einer staatlichen Mittelvergabe nicht ohne weiteres überwunden. Es ist somit durchaus möglich, dass eine bisher wenig erfolgreiche Einheit von einer bestimmten Ressourcenzuweisung mehr profitieren wird.

Eine sinnvolle Evaluation sollte den *marginalen Effekt einer Änderung der Mittel* erfassen. Was würde geschehen, wenn einer Institution oder einem Forscher mehr (oder weniger) Mittel zur Verfügung stehen? Diese Frage ist schwer zu beantworten, denn sie hängt von vielen institutionellen Bedingungen ab. Eine auf marginale Änderungen abstellende Evaluation ist wesentlich aufwendiger als die heute üblichen Ansätze. Das Verhältnis der Nutzen und Kosten von Evaluationen wird ungünstiger, weshalb es auch aus diesem Grund ratsam ist, sich ernsthaft mit den Alternativen zu Evaluationen auseinander zu setzen.

5. Institutionelle Änderungen als Alternative

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 (z.B. Kirchgässner, 2000, Frey 1990, 2001), insbesondere der „Institutionellen Ökonomik" (z.B. Richter und Furubotn, 1999, Erlei, Leschke und Sauerland, 1999) und der „Theorie der Wirtschaftspolitik" (Frey und Kirchgässner, 2002). Hier soll an einem konkreten Beispiel gezeigt werden, in welcher Weise eine bestimmte

institutionelle Ausgestaltung des Wissenschaftsbetriebs die heute üblichen Evaluationen zurückdrängen und teilweise sogar ersetzen können.

Werden Universitäten einem stärkeren Wettbewerb unterworfen, ist keine *staatliche* Evaluation mehr nötig. Die Studierenden zahlen kostendeckende Gebühren und wählen sich aber selbst diejenige Universität, die ihrer Ansicht nach die besten Leistungen bietet. Die Universitäten haben die Freiheit, sich diejenigen Studierenden auszuwählen, die ihre Anforderungen am besten erfüllen und die Reputation ihrer Hochschule steigern. Dieses System, bei dem sich sowohl die Nachfrager und Anbieter zwischen unterschiedlichen Möglichkeiten entscheiden können, ist nicht mehr auf eine Zuwendung der Mittel auf Grundlage einer (zentralen) Organisation angewiesen. Es gibt zwar Einschätzungen der Qualität der Hochschulen (sie werden häufig auch als „Evaluation" bezeichnet), diese werden jedoch privat angeboten und finanziert und dienen einem andern Zweck, nämlich die Studierenden zu informieren.

Der Zweck dieses Beispiels ist nicht, für ein Wettbewerbssystem privater Universitäten zu werben, sondern aufzuzeigen, dass es durchaus Alternativen zu den gängigen Evaluationen gibt.

6. Sorgfältige Personalauswahl als Alternative

Nachträgliche Evaluationen ganzer Universitäten, Disziplinen, Fakultäten, Fachbereichen, Instituten und Forschungsteams lassen sich weitgehend umgehen, wenn die Forschenden und Lehrenden sorgfältig ausgewählt werden. Diese Strategie setzt die Ressourcen *zukunftsorientiert* ein, indem überlegt wird, welche Personen vermutlich die zu bewältigenden Aufgaben bestmöglich erfüllen werden. Das Gewicht wird auf den Auswahlprozess gelegt (zu einem analogen Auswahlprozess in der Politik vgl. Cooter, 2002, Besley, 2005, 2006). Wenn eine Person einmal ernannt ist – zum Beispiel eine Professur für ein bestimmtes Wissensgebiet erhalten hat – wird ihr vertraut. Sie wird in Ruhe arbeiten gelassen. Aufgrund der sorgfältigen Auslese ist zu erwarten, dass sie die erwünschten Leistungen auch erbringen wird. Dabei ist mit einer erheblichen Varianz zu rechnen. Einige unter den ausgewählten Personen werden nicht mehr viel tun, 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. In den „Principles Governing Research at Harvard" (http://www.fas.harvard.edu/research/greybook/principles.html) wird genau dieses Vorgehen vertreten:

> „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."

Dauernde Evaluationen der Leistungen von Forschenden können hingegen in aller Regel ein bestimmtes Durchschnittsniveau sichern; die als Kontrolle

erlebten fortwährenden Beurteilungen führen entsprechend zu „normaler" Wissenschaft ohne Höchstleistungen. Es lässt sich schwer vorstellen, wie wirklich führende Forscher wie Einstein oder Planck in den Naturwissenschaften, Max Weber in der Soziologie, und Keynes oder Schumpeter in der Wirtschaftswissenschaft in einer durch dauernde Evaluationen geprägten Umgebung hätten florieren können. Sie wären nicht nur durch die notwendigen Rechtfertigungen ihrer Tätigkeit („Was haben Sie im letzten Halbjahr geforscht und veröffentlicht?") in ihrer eigentlichen Forschungstätigkeit aufgehalten worden, sondern hätten in einer Evaluation möglicherweise sogar schlecht abgeschnitten, weil sie die Prinzipien und Normen der „normalen" Wissenschaft (Kuhn, 1962) in Frage stellten und verwarfen. Viele Bahn brechende Beiträge zur Forschung wurden von den Zeitgenossen nicht verstanden und als lächerlich bezeichnet (siehe ausführlich Gillies, 2005).

Einer sorgfältige Auswahl und ein Vertrauen in den Willen und die Fähigkeit zur Leistung vermeidet eine nicht selten als Bevormundung aufgefasste Evaluation (vgl. zu den Auswirkungen von Vertrauen im Gegensatz zu Kontrolle Bohnet, Frey und Huck, 2001, Huang 2005). Die Bewertung der Leistung der Forschenden geschieht im dezentralen, autonomen und zuweilen langsamen Wissenschaftsprozess. Dieses System des grundsätzlichen Vertrauens in Personen hat in der Vergangenheit der deutschsprachigen Wissenschaft Weltgeltung verschafft. Es ist daher meines Erachtens verfehlt, die in den letzten Jahrzehnten nicht gerade überragende Leistung der Wissenschaft in deutschsprachigen Ländern auf die mangelnde Kontrolle mittels nachträglicher Evaluation zurückzuführen. Im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts ist im deutschen Sprachraum ähnlich wie noch heute in Harvard verfahren worden, und genau in dieser Zeit war die deutschsprachige Wissenschaft erstrangig. Häufig wird dagegen eingewandt, dass im deutschsprachigen Universitätssystem manche Professoren nichts mehr forschen und publizieren, sobald sie einmal eine lebenslange Anstellung erreicht haben. Dies gilt allerdings weitgehend auch für Angehörige anderer Universitätssysteme, wenn vielleicht auch dieses Versagen dort weniger sichtbar ist, weil die Betreffenden größere Anstrengungen unternehmen (müssen), dies zu verschleiern. Wesentlich ist jedoch, dass ein solches Versagen einzelner Personen die unvermeidlichen Kosten des Brillieren anderer Wissenschaftler ist, denen Raum für originelle und intensive Forschung eröffnet wird: das eine geht nicht ohne das andere.

7. Konklusion

Evaluationen im Sinne einer nachträglichen Bewertung der Leistung von Institutionen und Personen durch außen stehende Gutachtende vor allem zum Zwecke der Mittelzuweisung weisen einige „verborgene" Kosten auf. Im Vordergrund stehen schädliche Anreizverzerrungen, eine induzierte Verkrustung und ein verfehlter Entscheidungsansatz. Weil diese Kosten gewöhnlich nicht

berücksichtigt werden, werden Evaluationen zu oft und zu intensiv angewandt. Dies spricht nicht gegen Evaluationen an sich, wohl aber gegen deren heute festzustellende Dominanz und Allgegenwärtigkeit. Die Auffassung, die heutigen Evaluationen seien zwar mangelhaft, sollten aber einfach verbessert werden, wird nicht geteilt. Die hier vorgebrachten Einwände sind grundsätzlich und können nicht einfach mittels „besserer" Evaluationen überwunden werden. Die hier aufgeführten fundamentalen Probleme könnten sich noch verschlimmern.

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

Literaturverzeichnis

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

Amabile, T. (1998), How To Kill Creativity, *Harvard Business Review* 76(5), 76–87.

Backes-Gellner, U. und P. Moog (eds.) (2004), *Ökonomie der Evaluation von Schulen und Hochschulen*. Duncker und Humblot, Berlin.

Bedeian, A.G. (2003), The Manuscript Review Process: The Proper Roles of Authors, Referees, and Editors, *Journal of Management Inquiry* 12, 331–338.

Bénabou, R. und J. Tirole (2003), Intrinsic and Extrinsic Motivation, *Review of Economic Studies* 70, 489–520.

Besley, T. (2005), Political Selection, *Journal of Economic Perspectives* 19, 43–60.

Besley, T. (2006), *Principled Agents?* The Political Economy of Good Government. Oxford University Press, Oxford und New York.

Bohnet, I., B.S. Frey und S. Huck (2001), More Order With Less Law: On Contract Enforcement, Trust and Crowding, *American Political Science Review* 95, 131–144.

Bräuninger, M. und J. Haucap (2003), Reputation and Relevance of Economics Journals, *Kyklos* 56, 175–198.

Broadfoot, P. M. (1996), *Education, Assessment and Society*. Open University Press, Buckinham.

Brook, R. (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, T. und A. Stephan (2006), Do Eurozone Countries Cheat with their Budget Deficit Forecasts? *Kyklos* 59, 3–16.

Cameron, J., K.M. Banko und W.D. Pierce (2001), Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues, *The Behavior Analyst* 24, 1–44.

Chrystal, K.A. und P.D. Mizen (2003), Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy, in: P.D. Mizen (Hrsg.), *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart, Volume 1*. Edward Elgar, Cheltenham, U.K., 221–243.

Clark, W.C. und D. Cash (2001), From Science to Policy: Assessing the Assessment Process, KSF Faculty Research Working Paper No. RWP01–045, Kennedy School of Government, Harvard University.

Cooter, R.D. (2002), Who Gets to the Top in Democracy?: Elections as Filters, Working Paper No. 74, Berkeley Online Program in Law and Economics, University of California at Berkley.

Daily, C.M., D.R. Dalton und A.A. Cannella (2003), Introduction to Special Topic Forum. Corporate Governance: Decades of Dialogue and Data, *Academy of Management Review* 28, 371–382.

Daniel, H.-D. (1993), *Die Wächter der Wissenschaft*. Wiley-VCH, Weinheim.

Daniel, H.-D. und R. Fisch (Hrsg.) (1988), *Evaluation von Forschung: Methoden – Ergebnisse – Stellungsnahmen*. Universitätsverlag, Konstanz.

Day, P. und R. Klein (1990). *Age of Inspection. Inspecting the Inspectors*. Rowntree Foundation, London.

De Bruijn, H. (2002), *Managing Performance in the Public Sector*. Routledge, London und New York.

Deci, E.L., R. Koestner und R.M. Ryan (1999), A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation, *Psychological Bulletin* 125, 627–668.

Erlei, M., M. Leschke und D. Sauerland (1999), *Neue Institutionenökonomik*. Schäffer-Poeschel, Stuttgart.

Fehr, E. und S. Gächter (2002), Do Incentive Contracts Crowd Out Voluntary Cooperation? IEW Working Paper No. 34, Institute for Empirical Research in Economics, University of Zurich.

Figlio, D. und L. Getzler (2003), Accountability, Ability and Disability: Gaming the System, NBER Working Paper No 9307, National Bureau of Economic Research, Cambridge, MA.

Fischer, K. (1998), Evaluation der Evaluation, *Wissenschaftsmanagement* 5, 16–21.

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

Frey, B.S. (1992), Tertium Datur: Pricing, Regulation and Intrinsic Motivation, *Kyklos* 45, 161–184.

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

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

Frey, B.S. (2003), Publishing as Prostitution? Choosing between One's Own Ideas and Academic Success, *Public Choice* 116, 205–223.

Frey, B.S. und R. Jegen (2001), Motivation Crowding Theory, *Journal of Economic Surveys* 15, 589–611.

Frey, B.S. und G. Kirchgässner (2002), *Demokratische Wirtschaftspolitik*. Vahlen, München, 3. Auflage.

Frey, B.S. und M. Osterloh (2000), Pay for Performance – Immer empfehlenswert? *Zeitschrift für Führung und Organisation (ZFO)* 69, 64–69.

Frey, B.S. und M. Osterloh (2005), Yes, Managers Should Be Paid Like Bureaucrats. *Journal of Management Inquiry* 14, 96–111.

Frey, B.S. und M. Osterloh (2006), Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives, IEW Working Paper No. 302, Universität Zürich.

Gibbons, R. (1998), Incentives in Organizations, *Journal of Economic Perspectives* 12,115–132.
Gillies, D. (2005), Lessons from the History and Philosophy of Science regarding the Research Assessment Exercise. Paper read at the Royal Institute of Philosophy on 18 November 2005. (www.ucl.ac.uk/sts/gillies).
Gillies, D. (2006), Why Research Assessment Exercises Are a Bad Thing, *Post-Autistic Economics Review* 37, 2–9.
Goodhart, C.A. (1975), Monetary Relationships: A View from Threadneedle Street, in: *Papers in Monetary Economics, Volume I*. Reserve Bank of Australia.
Heckhausen, H. (1989), *Motivation und Handeln*. Springer, Berlin, 2. Auflage.
Holcombe, R.G. (2004), The National Research Council Ranking of Research Universities: Its Impact on Research in Economics, *Econ Journal Watch* 1, 498–514.
Holmstrom, B. und P. Milgrom (1991), Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design, *Journal of Law, Economics, and Organization* 7, 24–52.
Huang, F. (2005). To Trust or to Monitor: A Dynamic Analysis, Mimeo, School of Economics and Social Sciences, Singapore Management University.
Jordan, T.E. (1989), *Measurement and Evaluation in Higher Education: Issues and Illustrations*. Falmer Press, London.
Kirchgässner, G. (2000), *Homo Oeconomicus*. Mohr Siebeck, Tübingen, 2. Auflage.
Klostermeier, J. (1994), *Hochschul-Ranking auf dem Prüfstand: Ziele, Methoden und Möglichkeiten*. Interdisziplinäres Zentrum für Hochschuldidaktik der Universität Hamburg.
Kozar, G. (1999), *Hochschul-Evaluierung: Aspekte der Qualitätssicherung im tertiären Bildungsbereich*. WUF, Wien.
Kuhn, T.S. (1962), *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.
Lindsey, D. (1989), Using Citation Counts as a Measure of Quality in Science: Measuring What's Measurable Rather Than What's Valid, *Scientometrics* 15, 189–203.
Lucas, R.E. (1976), Econometric Policy Evaluation: A Critique, *Carnegie-Rochester Conference Series on Public Policy* 1, 19–46.
Max Planck Gesellschaft (2002), *Innovative Structures in Basic Decision Research*, Ringberg Symposium, 4.-7. Oktober 2000 in München.
McCabe, D.L., L.K. Trevino und K.D. Butterfield (1996), Cheating in Academic Institutions: A Decade of Research, *Ethics and Behavior* 11, 219–232.
Neave, G. (1988), On the Cultivation of Quality, Efficiency and Enterprise: An Overview of Recent Trends in Higher Education in Western Europe, 1986–1988, *European Journal of Education* 23, 7–23.
Osterloh, M. und B.S. Frey (2005), Shareholders Should Welcome Employees as Directors, IEW Working Paper No. 228, Institute for Empirical Research in Economics, University of Zurich.
Power, M. (1994). *The Audit Explosion*. Demos, London.
Power, M. (1997), *The Audit Society. Ritual of Verification.* Oxford University Press, Oxford.
Richter, R. und E. Furubotn (1999), *Neue Institutionenökonomik*. Mohr Siebeck, Tübingen.
Röbbecke, M. und D. Simon (1999), Zwischen Reputation und Markt – Ziele, Verfahren und Instrumente von (Selbst)Evaluationen außeruniversitärer, öffentlicher Forschungseinrichtungen, WZB-Discussion Paper P. 99–601, Wissenschaftszentrum Berlin.

Röbbecke, M. und D. Simon (2001), Assessment of the Evaluation of Leibniz-Institutes – External Evaluation and Self-Evaluation, in: P. Shapira und S. Kuhlmann (Hrsg.), *Proceedings from the 2000 US-EU Workshops on Learning from Science and Technology Policy Evaluation*. Bad Herrenalb, 16–23.

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

Russon, C. und K. Russon (Hrsg.) (2000), *The Annotated Bibliography of International Programme Evaluation*. Kluwer, Dordrecht.

Schulze, G. und B. Frank (2003), Deterrence versus Intrinsic Motivation: Experimental Evidence on the Determinants of Corruptibility, *Economics of Governance* 4, 143–160.

Starbuck, W.H. (2004), Methodological Challenges Posed by Measures of Performance, *Journal of Management and Governance* 8, 337–343.

Stockmann, R. (ed.) (2004), *Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder*. Leske und Budrich, Opladen, 2. Auflage.

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

Weingart, P. (2005), Impact of Bibliometrics upon the Science System: Inadvertent Consequences?, *Scientometrics* 62, 117–1.

Abstract: *In the sciences the outside evaluation of past performances of universities, faculties, departments, research groups and of individuals has become more and more frequent, nearly incessant. It could be said that the sciences are afflicted with "Evaluitis", a creeping and widespread illness. Besides the obvious costs that arise for those being evaluated and for those doing the evaluation there are additional costs that weigh heavily but are usually disregarded: incentives are distorted systematically and ossification is promoted. Furthermore, the whole decision approach is wrongly conceived. For these reasons there are too many and too thorough evaluations. A useful alternative is an appropriate design of institutions guiding incentives and a careful selection of persons – who thereafter should be free to pursue their tasks.*