CREMA

Center for Research in Economics, Management and the Arts

Evaluitis – Eine neue Krankheit

Bruno S. Frey

Working Paper No. 2006 - 18

CREMA Gellertstrasse 24 CH - 4052 Basel www.crema-research.ch

WZB-Konferenz, 1. – 3. Juni 2006
"Qualitätssicherung von Wissenschaft im Wandel"

EVALUITIS – EINE NEUE KRANKHEIT

von

Bruno S. Frey*

Universität Zürich
und
CREMA- Center for Research in Economics, Management and the Arts

(9 Mai 2006)

Abstract

"Evaluitis" - i.e. ex post assessments of organizations and persons - has become a rapidly spreading disease. In addition to the well-known costs imposed on evaluees and evaluators, additional significant costs are commonly disregarded: incentives are distorted, ossification is induced and the decision approach is wrongly conceived. As a result, evaluations are used too often and too intensively. A viable and often superior alternative to evaluations is a careful selection of persons and afterwards leaving them to pursue their assigned tasks.

Keywords: Evaluation, Performance, Selection, Research, Incentives

JEL Classification: D23, D61, H43, M40, M5

* Institut für Empirische Wirtschaftsforschung der Universität Zürich, Blümlisalpstrasse 10, CH-8006 Zürich, Schweiz. E-Mail: bsfrey@iew.unizh.ch.
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.

In den letzten Jahren ist eine neue, sich fieberhaft ausbreitende Krankheit ausgebrochen: Jedes und alles wird unablässig evaluiert.

Unter „Evaluation“ wird hier eine *nachträgliche Einschätzung der Leistung einer Organisation oder Person durch von aussen kommende Experten verstanden*[1]. Diese Arbeit konzentriert sich auf Evaluationen im staatlichen Auftrag, die insbesondere helfen sollen, die geeignete Zuteilung finanzieller Mittel zu unterstützen.

Die Krankheit *Evaluitis* hat ganz besonders auch die Wissenschaft befallen. Heute werden in immer kürzeren Abständen ganze Universitäten, Fakultäten, Fachbereiche, Institute, Forschungsgruppen und einzelne Forschende begutachtet. Evaluationen und daraus abgeleitete Rankings sind heute 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.

Dieser Beitrag möchte auf einige wenig diskutierte, verborgene und damit gewöhnlich vernachlässigte Kosten von Evaluationen aufmerksam machen. In soweit diese Kosten nicht berücksichtigt werden, wenn darüber entschieden wird, ob eine Evaluation durchgeführt werden sollte (falls darüber überhaupt noch entschieden wird), wird der Nettonutzen dieses Instrumentes *systematisch* überschätzt. In diesem Falle werden *zu viele* und *zu intensive* Evaluationen durchgeführt als gesellschaftlich sinnvoll wäre. Insofern lässt sich „Evaluitis“ als eine Krankheit bezeichnen. Ich möchte jedoch deutlich machen, dass dies *kein* Argument gegen Evaluationen an sich ist; in manchen Fällen sind sie notwendig und sinnvoll. Allerdings wird nicht die Auffassung geteilt, die heutigen Evaluationen seien zwar mangelhaft, sollten aber einfach verbessert werden. Die hier vorgebrachten Einwände sind grundsätzlich und lassen sich nicht einfach beseitigen, indem die Evaluationen differenzierter werden. Verbesserte, und damit intensivere Evaluationen können möglicherweise die hier aufgeführten fundamentalen Probleme sogar noch verschlimmern.

---

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

Zum Thema Evaluation existiert eine riesige Literatur, die hier nicht wiederholt wird[2]. In diesem Beitrag wird somit nicht auf die sattsam bekannten Kosten in Form von Material und Zeit auf Seiten der Evaluierer und der Evaluierten eingegangen[3]. Ebenso wenig wird auf die offensichtlichen Nutzen von Evaluationen für die Entscheidungsbildung eingegangen. Im Zentrum stehen somit gerade bei der Anwendung in der Praxis vernachlässigte Aspekte[4]. 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[5]. 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.
Im ersten Teil werden die vernachlässigten Kosten von Evaluationen diskutiert. Abschnitt I befasst sich mit der durch Evaluationen verursachten Anreizverzerrung bei den Evaluierten; Abschnitt II mit der induzierten Verkrustung und Abschnitt III mit dem verfehlten Entscheidungsansatz und damit dem geringen Nutzen für die Entscheidungsbildung. Im zweiten Teil werden die Alternativen zu den Evaluationen behandelt. Abschnitt IV argumentiert, dass ein gewünschtes Verhalten auch mittels institutionellen Änderungen und einer sorgfältigen Personenauslese erzielt werden kann. Im letzten Teil werden Folgerungen gezogen.

I. Evaluation verzerrt Anreize

Das Instrument der Evaluation verändert das Verhalten der davon betroffenen Personen in systematischer und auch unbeabsichtigter Weise. Es darf somit nicht

---

[2] 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, Cash und Clark 2001, Bräuninger und Haukap 2003, und die Zeitschriften *Research Evaluation* und *Scientometrics.*

[3] 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 pollsters' questions".

[4] Es wird natürlich nicht behauptet, dass diese nirgendwo erwähnt würden, wohl aber dass sie kaum oder gar nicht beachtet werden.

[5] So etwa bei Holcombe 2004, Starbuck 2004, Royal Netherlands Academy of Arts and Sciences 2005, Weingart 2005.

davon ausgegangen werden, dass sich Individuen (und entsprechend auch Institutionen) infolge einer Evaluation ihr Verhalten in der von den Evaluierten gewünschten Weise verändern, vor allem zielorientierter und effizienter arbeiten. Vielmehr werden auch unerwünschte Verzerrungen im Verhalten ausgelöst: (A) Eine Konzentration auf das, was gemessen wird; (B) Eine Verdrängung intrinsischer Arbeitsanreize, wodurch vor allem die Originalität betroffen wird; und (C) Eine Manipulation der Kennziffern.

A. *Was nicht gemessen wird, zählt nicht (mehr).*

Das Phänomen des „Multi-tasking" ist in der Wirtschaftswissenschaft seit einigen Jahren intensiv diskutiert worden[6]: Die Vorgesetzten (Prinzipale) legen die Massstäbe fest, mit denen die Leistung einer Institution oder einer Person beurteilt werden. Für keine Tätigkeit – ausser möglicherweise einfachster Fliessbandtätigkeit – lässen sich jedoch *alle* relevanten Aspekte definieren und messen. Jede Person hat deshalb die Tendenz, oder wird sogar gezwungen, sich ausschliesslich nach den gemessenenen 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.

In der Wissenschaft hat das multiple-tasking Problem besonders starke Auswirkungen. So wird heute fast überall die Anwerbung von Drittmitteln als Leistungskriterium verwendet (so etwa bei den Organisationen in der Leibniz-Gesellschaft, wozu auch die Max Planck Gesellschaften gehören). Dass damit weder der Sinn noch die Produktivität der damit finanzierten Forschung erfasst wird, ist augenscheinlich. Dieses Mass ist jedoch gängig, weil Geldströme besonders leicht messbar sind. Wenn aber eine wissenschaftliche Einheit damit 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 (1989:200) „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,

[6] Vgl. z.B. Holmstrom und Milgrom 1991, Gibbons 1998, Daily, Dalton und Cenella 2003, Suvorov und van de Ven 2006.

allgemeinbildende Vorträge, Beratungstätigkeit, 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. Anstelle weniger Gewicht auf derartige Evaluationen zu legen wird vielmehr versucht, die entsprechenden Aspekte ebenfalls quantitativ zu erfassen. Dies wird jedoch *nie* im vollen Umfang möglich sein. Das multiple 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: 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 veranlasst, sich solchen Forschungsfragen zu widmen, die gerade Mode sind und wo sie deshalb erwarten können, viel zitiert zu werden. In vielen Disziplinen dürfte damit die angewandte Forschung benachteiligt werden.

B. *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 (eine umfassende Metastudie ist 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 2004) empirisch anhand von Felduntersuchungen bestätigt worden (eine Übersicht geben Frey und Jegen 2001). Der Verdrängungseffekt besagt, dass infolge der als kontrollierend empfundenen Evaluation die intrinsische Arbeitsmotivation abnimmt und die extrinsisch bestimmten Anreize an Gewicht gewinnen. Die Gesamtleistung vermindert sich nicht notwendigerweise, sondern steigt sogar für manche Evaluierte. Dies entspricht dem Ergebnis des britischen „ Research Assessment Exercise“. Gemäss Brook (2002: 176) „ .. we can safely say that the average activity has increased“ – zumindest in der

Evaluation erfassten Dimension. Es muss jedoch bezweifelt werden, ob die Auswirkungen auf die Qualität und Originalität der Forschung günstig waren. Wie Amabile (1996, 1998) gezeigt hat, ist die intrinsische Motivation für innovative wissenschaftliche Arbeit von entscheidender Bedeutung. Hinzu kommt, dass gerade bahnbrechende Forschung gegen den Konsens der Evaluierenden verstösst und deshalb gering geschätzt wird. Historische Untersuchungen (Fischer 1996, Gillies 2006) zeigen, dass viele besonders wichtige Forschungsergebnisse dem Zeitgeist (im Sinne der „normal science" von Kuhn 1962) widersprachen und deshalb in einer Evaluation schlecht 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 immer mehr als kontrollierend empfunden und die intrinsische Forschungsmotivation wird immer mehr verdrängt.

Der Verdrängungseffekt ist quantitativ schwer zu fassen und wird deshalb leicht vernachlässigt. Im Umfang, in dem dies der Fall ist, wird zuviel, zu häufig und zu intensiv evaluiert.

C. Manipulation der Leistungskriterien

Wenn ein Indikator für die eigene Position wichtig wird, wird ein starker Anreiz ausgeübt, diesen Indikator zu seinen eigenen Gunsten 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. Chrystal and Mizen 2003, Brück und Stephan 2006). Er gilt 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 2000, 2005, Frey und Osterloh 2005).

Eine derartige Manipulation hat sich auch in der Wissenschaft verbreitet, seit im Zuge von Evaluationen die Forschungsleistung anhand der Zahl der Publikationen und Zitierungen gemessen wird. So werden vorübergehend Wissenschaftler mit entsprechenden Leistungsausweisen an eine Universität verpflichtet, damit diese in Evaluationen gut abschneidet. Nicht selten haben diese Forscher nur eine lose, oder sogar keine Beziehung zu diesen Universitäten und deren Forschungsleistung wird von mehreren Universitäten gleichzeitig benützt. Für die Wissenschaftskultur 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).

II. Induzierte Verkrustung

Evaluationen bewirken „lock-in"-Effekte 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.

A. 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 all zu 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 besserer Einsicht an der Evaluation

beteiligen. Sie tun sogar gut daran, begeistert mit zu machen. Nach aussen wird dadurch der Anschein erweckt, die Evaluierten seien von den Vorzügen einer Evaluation überzeugt. Damit wird ein Einverständnis vorgetäuscht, das in Wirklichkeit nicht vorhanden ist. Gleichzeitig wird einer zynischen Haltung zur Wissenschaft und deren Ergebnissen Vorschub geleistet.

B. 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 abhängt. Sie sind deshalb bestrebt, Evaluationen auf immer weitere Bereiche auszudehnen, zu intensivieren und in immer kürzeren Abständen durchzuführen. Am vorteilhaftesten ist eine dauernde Evaluation, wofür sich viele Argumente vorbringen lassen. Hingegen werden die negativen Aspekte von Evaluationen, wie etwa die im letzten Abschnitt aufgeführten wenig sichtbaren Kosten, heruntergespielt. Dieses aktive Lobbying oder „Rent seeking“ trägt zur Ausweitung der Evaluationen bei. Gleichzeitig wird den im zweiten Teil dieses Aufsatzes genannten Alternativen zur Evaluation wenig oder gar keinen Raum gegeben.

III. Geringer Nutzen von Evaluationen für Entscheidungen

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. Es fällt schwer zu sehen, warum diese zusätzliche Information nicht so nützlich ist, wie sie auf den ersten Blick erscheint. Dazu sind zwei Gründe massgeblich.

A. Wenig Informationsgewinn

Gerade in der Wissenschaft ist in der „Scientific Community“ häufig sehr wohl bekannt, welche Institutionen und Personen gute Forschung betreiben. Bestätigt die Evaluation dieses Ergebnis, ist wenig oder nichts gewonnen. Kommt sie hingegen zu einem anderen Ergebnis, wird dieses 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. Wie in Abschnitt B gezeigt werden wird, stehen dafür gute Argumente zur Verfügung. Auf jeden Fall kann nicht davon ausgegangen werden, dass negative Evaluationsergebnisse grosse Auswirkungen haben. Oft sind sie nur symbolischer Natur.

In manchen Fällen besteht in der Scientific Community keine Einigkeit über die Qualität eines Forschenden oder einer Forschungsorganisation. In diesem Fall verbessert eine Evaluation die Information. Typischerweise dürften jedoch die staatlichen Entscheidungen keine besonders überraschenden Ergebnisse liefern, sondern mehr oder weniger die Position im Mittelfeld bestätigen. Dies bedeutet wohl auch, dass die Mittelzuweisung zumindest im Vergleich zu anderen Institutionen wenig verändert wird. Der (hohe) Evaluationsaufwand lohnt sich deshalb nicht notwendigerweise.

B. Für Entscheidungen irrelevante Information

Evaluationen suchen in aller Regel das bestehende Leistungsniveau anhand einer grossen Zahl 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.

Eine Evaluation sollte idealerweise 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 äusserst schwierig zu beantworten, denn sie hängt von einer grossen Zahl von Bedingungen ab. Ausserdem

bleibt auch hier offen, wie die Ergebnisse einer derartigen Evaluation in der politischen Auseinandersetzung aufgenommen würden. Eine auf marginale Änderungen abstellende Evaluation ist wesentlich aufwendiger als die heute üblichen Ansätze, was das Verhältnis der Nutzen und Kosten von Evaluationen beeinträchtigt. Aus diesem Grund ist ratsam, sich ernsthaft mit den Alternativen zu Evaluationen zu beschäftigen.

IV. Die Alternative institutioneller Änderungen

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 „Instutionellen Ökonomik"(z.B. Richter und Furubotn 1999, Erlei, Leschke und Sauerland 1999) und der „Theorie der Wirtschaftspolitik" (Frey und Kirchgässner 2002). Diese empirisch in Hunderten von Studien nachgewiesenen Wirkungen brauchen an dieser Stelle nicht weiter ausgeführt werden. Vielmehr 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 Studiengebühren, 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.

V. Die Alternative sorgfältiger Personalauswahl

Die heute üblich gewordenen nachträgliche Evaluationen ganzer Universitäten, Fakultäten, Fachbereichen, Instituten und Forschungsteams lässt sich zu einem guten Teil umgehen, wenn die Forschenden und Lehrenden sorgfältig ausgewählt werden. Diese Strategie setzt die Ressourcen *zukunftsorientiert* ein, indem Personen bestmöglich mit den zu bewältigenden Aufgaben betraut werden. Das Gewicht wird auf den Auswahlprozess gelegt[7]. 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 und es wird aufgrund der sorgfältigen Auslese erwartet, dass sie die erwarteten 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. Eine derartige Organisation der Wissenschaft wird entschieden von James Bryan Conant, dem bedeutenden Präsidenten der Harvard Universität, vertreten:

> „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, zitiert 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“

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 Höhenleistungen. Diese Situation wird dadurch verstärkt, dass sich - wie oben ausgeführt - kaum jemand oder gar niemand anstehenden Evaluationen entziehen kann. Es lässt sich schwer vorstellen, wie wirklich führende Forscher wie Einstein oder Planck in den Naturwissenschaften, und Keynes oder Hicks in meiner eigenen Wissenschaft in einer durch dauernde Evaluationen geprägten Umgebung florieren

---

[7] Zu einem analogen Auswahlprozess in der Politik vgl. Cooter 2002, Besley 2005.

könnten. Sie würden nicht nur durch die notwendigen Rechtfertigungen ihrer Tätigkeit („Was haben Sie im letzten Halbjahr geforscht und veröffentlicht?“) in ihrer eigentlichen Forschungstätigkeit aufgehalten, sondern würden in einer Evaluation möglicherweise sogar schlecht abschneiden, weil sie die Prinzipien und Normen der „normalen“ Wissenschaft (Kuhn 1962) in Frage stellen und verwerfen. Viele bahnbrechende Beiträge zur Forschung wurden von den Zeitgenossen nicht verstanden und als lächerlich bezeichnet. Freges im Jahre 1897 veröffentlichte innovative mathematische Theorie wurde in fünf von sechs Besprechungen extrem herablassend eingeschätzt, dessen Bedeutung wurde erst zwanzig Jahre später allmählich erfasst (u.a. durch Bertrand Russell) und ist erst in den 1950er Jahren allmählich anerkannt worden. Die Forschungen von Semmelweis zur Antisepsis (1847) wurden erst rund zwanzig Jahre später akzeptiert und die fundamentalen astrologischen Einsichten von Copernicus (1473-1543) wurden während dessen Lebenszeit und bis 50 oder 60 Jahre später von anderen Astronomen als absurd angesehen (siehe ausführlich Gillies 2005).

Die Alternative einer sorgfältigen Auswahl und dann eines Vertrauens in den Willen und die Fähigkeit zur Leistung vermeidet eine nicht selten als Bevormundung aufgefasste Evaluation[8]. Die Bewertung der Leistung der Forschenden geschieht im dezentralen, autonomen und zuweilen langsamen Wissenschaftsprozess. Es sollte nicht vergessen werden, dass dieses System in der Vergangenheit der deutschsprachigen Wissenschaft eine Weltgeltung verschafft hat. Wenn es durch das System dauernder Evaluationen ersetzt werden soll, müssen überzeugende Argumente vorgebracht werden, warum es nicht mehr wirksam sein soll.

VI. 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. 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. Wie betont wurde, spricht dies nicht gegen Evaluationen an sich, wohl aber gegen deren heute festzustellende

[8] Die Auswirkungen von Vertrauen im Gegensatz zu Kontrolle wird analysiert bei Bohnet, Frey und Huck 2001, Huang 2005.

Dominanz und Allgegenwärtigkeit. Ebenso wird nicht die Auffassung geteilt, die heutigen Evaluationen seien zwar mangelhaft, sollten aber einfach verbessert werden. Die hier vorgebrachten Einwände sind grundsätzlich und können nicht einfach beseitigt werden, indem die Evaluationen differenzierter werden. Es ist sogar denkbar, dass verbesserte, und damit intensivere Evaluationen die hier aufgeführten fundamentalen Probleme noch verschlimmern.

Die häufig vorgebrachten 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.

Literatur

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

Amabile, Teresa (1998). How To Kill Creativity. *Harvard Business Review,* Sep.-Oct. 76(5): 76-87.

Backes-Gellner, Uschi und Petra Moog (eds) (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. *Journal of Management Inquiry*, 12: 331-338.

Bénabou, Roland und Jean Tirole (2003). Intrinsic and Extrinsic Motivation. *Review of Economic Studies* 70 3: 489-520.

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

Bohnet, Iris, Bruno S. Frey und Steffen Huck (2001). More Order With Less Law: On Contract Enforcement, Trust and Crowding. *American Political Science Review,* Vol. 95, No. 1: 131-144.

Bräuninger, Michael und Justus Haukap (2003). Reputation and Relevance of Economics Journals. *Kyklos* 56: 175-198.

Broadfoot, P.M., (1996) *Education, Assessment and Society.* Buckinham: 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 und Andreas Stephan (2006). Do Eurozone Countries Cheat with their Budget Deficit Forecasts? *Kyklos* 59: 3-16.

Cameron, J., Banko, K. M. und W.D. Pierce (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst,* 24: 1-44.

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

Clark, Burton R. (1983). *The Higher Education System. Academic Organization in Cross-National Perspective.* Berkeley: University of California Press.

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

Cooter, Robert D. (2002). Who gets to the Top in Democracy?: Elections as Filters. Working Paper Series no. 74, Berkeley Online Program in Law and Economics.

Daily, Catherine M., Dan R. Dalton und Albert. A. Cannella (2003). Introduction to Special Topic Forum. Corporate Governance: Decades of Dialogue and Data. *Academy of Management Review* 28(3): 371-382.

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

Daniel, Hans-Dieter und Rudolf Fisch (eds) (1988). *Evaluation von Forschung: Methoden-Ergebnisse-Stellungsnahmen.* Konstanz: Universitätsverlag.

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

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

Deci, Edward L., Richard Koestner und Richard M. Ryan (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125(6): 627-668.

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

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

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

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

Fischer, Klaus (1998). Evaluation der Evaluation. *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. *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. *Public Choice*. 116: 205-223

Frey, Bruno S. (2005). Knight Fever. Towards an Economics of Awards. IEW Working Paper No. 239, Institute for Empirical Research in Economics, University of Zurich.

Frey, Bruno S. und Reto Jegen (2001). Motivation Crowding Theory. *Journal of Economic Surveys* 15(5): 589 - 611.

Frey und Kirchgässner (2002). *Demokratische Wirtschaftspolitik.* München: Vahlen, 3. Auflage.

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

Frey, Bruno S. und Margit Osterloh (2005). Yes, Managers Should Be Paid Like Bureaucrats. *Journal of Management Inquiry,* 14: 96-111.

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

Fröhlich, Klaus (2006). "Informed Peer Review" - Ausgleich der Fehler und Verzerrungen? In HRK (Hochschulrektorenkonferenz) (ed), *Von der Qualitätssicherung der Lehre zur Qualitätsentwicklung der Hochschulsteuerung.* Bonn: 193-204.

Gibbons, Robert (1998). Incentives in Organizations. *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. *post-autistic economics review* 37: 2-9.

Goodchild, L.F., C.D. Lovell. E.R. Hines and J.L. Gill (1997). *Public Policy and Higher Education.* Needham Heights.

Halcombe, Randall G. (2004). The National Research Council Ranking of Research Universities: Its Impact on Research in Economics. *Econ Journal Watch* 1: 498-514.

Heckhausen, Heinz (1989). *Motivation und Handeln*. Zweite, völlig überarbeitete und ergänzte Aufl. Berlin etc.: Springer.

Holmstrom, Bengt und Paul Milgrom (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization* 7(2): 24-52.

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.

Kieser, Alfred (1998). Going Dutch - Was lehren niederländische Erfahrungen mit der Evaluation universitärer Forschung? *DBW* 58: 208-224.

Kirchgässner, Gebhard (2000). *Homo Oeconomicus.* Tübingen: Siebeck, 2. Auflage.

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

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.

Latham, Gary P., Joan Almost, Sara Mann and Cecilia Moore (2005). New Developments in Performance Management *Organiazional Dynamics* 34: 77-87.

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

McCabe, D.L., L. Trevino und K. Butterfield (1996). Cheating in Academic Institutions: A Decade of Research. *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. *European Journal of Education*, 23 (1-2): 7-23.

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

Pasternack, Peter (2000). Besoldete Qualität? Qualitätsbewertung und leistungsgerechte Besoldung. *Wissenschaftsmanagement* 4: 8-13.

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 und Erik Furubotn (1999). *Neue Institutionenökonomik* . Tübingen: Siebeck.

Röbbecke, Martina und Dagmar Simon (1999). *Zwischen Reputation und Markt – Ziele, Verfahren und Instrumente von (Selbst)Evaluationen ausseruniversitärer, öffentlicher Forschungseinrichtungen.* WZB-Discussion Paper, P 99-601, 83 S.

Röbbecke, Martina und Dagmar Simon (2001). Assessment of the Evaluation of Leibniz-Institutes – External Evaluation and Self-Evaluation. In: Philip Shapira and Stefan Kuhlmann (eds.), *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 und Karen Russon (eds) (2000). *The Annotated Bibliography of International Programme Evaluation.* Dordrecht: Kluwer.

Sanders, James R. (ed) (2000). *Handbuch der Evaluationsstandards.* Opladen: Leske.

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

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

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

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

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