MARTIN FROMM

Methodologische Fragen zur Evaluation von Lehr-/Lernprozessen in der Hochschule: Wissenschaftsforschung im Auftrag der Medien am Beispiel der SPIEGEL-Studien

1. Hochschulevaluation im Auftrag der Medien

Bereits seit einigen Jahren ist die Evaluation universitärer Lehre und Forschung nicht mehr nur Angelegenheit der Universitäten selbst oder der Bildungspolitik. Vielmehr haben hier die Medien, insbesondere die Printmedien, zunehmend eine aktive Rolle übernommen, indem sie nicht nur berichten und kommentieren, sondern selbst initiativ werden und Evaluationsstudien in Auftrag geben. Der Studie des SPIEGELs 1989/90 folgte 1992/93 eine weitere des SPIEGELs, im gleichen Jahr Studien von STERN und FOCUS und 1998 nun wieder eine des SPIEGELs. Weitere Auftragsstudien zur Situation an den Hochschulen anderen Zeitschriften wären zu ergänzen (z.B. die Umfrage „Student 95" in „DIE ZEIT").

Die Ergebnisse dieser Studien sollen hier ebenso wenig interessieren, wie die mutmaßliche Wirkung, die diese auf das öffentliche Bild der Universitäten haben, welche bildungspolitischen Einflüsse von derartigen Studien zu erwarten sind usw. Die folgenden Überlegungen konzentrieren sich vielmehr auf die Frage, wie diese Studien Einfluss auf den empirischen Evaluationsprozess selbst und allgemeiner empirische Forschung nehmen, welche Vorstellungen sie davon erzeugen oder beeinflussen, was Evaluation leisten kann und soll, was gute von schlechten Evaluationskonzepten unterscheidet, welche Erhebungs-, Analyse- und Darstellungsverfahren geeignet/ungeeignet sind, usw. Exemplarisch soll das an den SPIEGEL-Studien untersucht werden, die nicht nur durch Beteiligung von Wissenschaftlern – zunächst NEIDHARDT, dann HORNBOSTEL und DANIEL - den Anspruch auf methodisch kompetente Abwicklung und Seriosität der Studien erheben, sondern auch Gegenstand ausgedehnter forschungsmethodischer Diskussionen gewesen sind (vgl. z. B. MOHLER 1995). Da diese Studien zudem modifiziert fortgesetzt worden sind, bietet sich hier auch eine Möglichkeit zu überprüfen, ob und wie sich methodische Kritik in veränderten Vorgehensweisen niederschlägt.

2. Die SPIEGEL-Studien: Fragestellung und Aufbau

Die SPIEGEL-Studien streben einen Leistungsvergleich ausgewählter Studiengänge an verschiedenen Universitätsstandorten an, um einmal im Wettbewerb der Hochschulen Anhaltspunkte für eine gezielte Verbesserung der Studienbedingungen zu

liefern, zum anderen, um Studierenden eine Orientierung zu bieten, was sie an welchen Universitäten erwartet.

Basis des Vergleichs sind Studentenurteile über die Studienbedingungen in einem Studienfach an der Universität, an der sie studieren. Beurteilt wird dabei z. B. die materielle Ausstattung (z. B. Anzahl von Arbeitsplätzen, Verfügbarkeit von Fachliteratur) und die Qualität der Lehre (z. B. Vorbereitung der Dozenten, Verständlichkeit). Erfasst werden diese Urteile als Antworten auf geschlossene Fragen, z. B. „Bemühen sich sehr viele (Skalenwert 1) oder nur wenige (Skalenwert 6) Dozenten darum, die Studierenden auf das Examen vorzubereiten?"

1989/90 wurden insgesamt 6000 Studierende an deutschen Universitäten befragt (12 pro Fachbereich), 1992/93 über 10000 (18 pro Fachbereich). 1998 wurde die Untersuchung einerseits auf Europa ausgedehnt, andererseits die Anzahl der Universitäten in Deutschland und die Anzahl der Fachbereiche reduziert, so dass nun insgesamt 7434 Studenten (ca. 50 pro Fachbereich) untersucht wurden.

3. Wissenschaftliche Kritik an den SPIEGEL-Studien

Insbesondere die Studie von 1989/90 war als Initialstudie Gegenstand ausgedehnter öffentlicher Diskussionen. Zur öffentlichen Diskussion etwa der Frage, wie man denn etwas so komplexes wie Bildung oder Studienqualität sinnvoll messen könne, kam eine lebhafte methodologische Diskussion, die sich mit der Frage beschäftigte, wie weit diese Studie und dann die von 1992/93 nach gängigen forschungsmethodischen Gütestandards zu beurteilen sei. Ich konzentriere mich auf diese wissenschaftsmethodologische Diskussion und hier auf die Kritikpunkte, die erstens nicht nur Details oder die methodische Eleganz betreffen und sich zweitens auf Bewertungskriterien stützen, die als weitestgehend akzeptiert gelten können. Kritisiert wird u. a.:

- Größe und Ziehung der untersuchten Stichproben lassen die Repräsentativität der Ergebnisse als höchst zweifelhaft erscheinen.
- Es werden Daten auf einer Skala zusammengefasst und damit relationiert, die in keiner erkennbaren Beziehung zueinander stehen.
- Es werden Rangunterscheidungen vorgenommen, die messtheoretisch nicht vorgenommen werden können.

Repräsentativität

Hier sind zwei Aspekte zu unterscheiden: Um von einer Stichprobe auf eine Grundgesamtheit schließen zu können, ist einmal wichtig, dass man die richtigen Versuchspersonen befragt, zweitens, dass man genug von ihnen befragt. Beim ersten Aspekt geht es darum, dass die untersuchte Stichprobe in ihren wesentlichen Merkmalen der Grundgesamtheit entspricht, über die etwas ausgesagt werden soll, quasi also ein verkleinertes Abbild dieser Grundgesamtheit darstellt. Beim zweiten Aspekt geht

es dagegen darum, mit welcher Wahrscheinlichkeit man bei den zu erwartenden Messfehlern von der Stichprobe auf die Grundgesamtheit schließen kann.

Erkennbar hängen dabei beide Aspekte zusammen: Wenn zwar die richtigen Versuchspersonen, aber zu wenige von ihnen befragt werden, sind Schlüsse auf die Grundgesamtheit messtheoretisch extrem unsicher, und wenn man viele Versuchspersonen, aber eben die falschen, befragt, ist der Schluss inhaltlich unsinnig.

Im Fall der SPIEGEL-Studien von 1989 und 1993 wurde die Repräsentativität der Stichproben in beiden Aspekten in Frage gestellt (vgl. z.B. KRIZ 1995, S. 20), wobei die geringen Stichprobengrößen (12 bzw. 18 Studenten pro Fach) bekannt waren, Mängel der Stichprobenziehung – z. B. bevorzugte Befragung von leicht erreichbaren Studenten – nur vermutet wurden. Zusammengefasst: Es wurden die falschen Studenten befragt, und davon auch noch zu wenig.

Relationierung von Daten

Ein weiterer Kritikpunkt setzt bei dem Versuch an, einen Vergleich zwischen Hochschulen, bzw. einzelnen Fächern an verschiedenen Hochschulen herzustellen. Voraussetzung dafür ist erkennbar, dass die Daten, die man an den einzelnen Hochschulen gewinnt, sinnvoll zueinander in Beziehung gesetzt werden können.

Im Fall der SPIEGEL-Studien werden den Studierenden jeweils Urteile über ihre eigene Universität bzw. ihr eigenes Studienfach abverlangt. Das setzt natürlich voraus, dass man ihnen zutraut, die angesprochenen Gegebenheiten an ihrer Universität hinreichend zu beurteilen. KRIZ (1995, S. 19) fragt z. B., ob man denn sinnvoll von den Studierenden eine Beurteilung des Bestandes an 'wichtigen' Büchern erwarten könne, der ja schließlich nach Vorgaben ihrer Dozenten zustande gekommen sei. Kritischer ist dann aber die Frage, „ob ein Hamburger Student den Bücherbestand in Hamburg mit dem in München (und allen anderen Unis) vergleichen kann – ohne jemals dort gewesen zu sein" (ebd.).

Die Kritik besagt hier, dass die Relationierung der Urteile, die Studierende an den verschiedenen Universitäten fällen, keine erkennbare Grundlage habe, entsprechend gänzlich unklar bleibe, was die artifiziell geschaffene Skalierung eigentlich repräsentiere.

Ranking

Erklärter Zweck der Studien ist auch, eine möglichst klare Antwort auf die Frage zu geben „Welche Uni ist die beste?", im „Uni-Test Europa" von 1998: „Welches Land hat die besten Unis?" (DER SPIEGEL 1998, S. 95). Dazu müssen dann die Ergebnisse der einzelnen Universitäten über mehrere Fächer oder Urteilsdimensionen hinweg verglichen werden.

Wenn man von den Grundlagen des Vergleichs und der inhaltlichen Sinnhaftigkeit einmal absieht, ist an dieser Stelle ein messtheoretischer Einwand, zum Zweck des Rankings würden Messfehler ignoriert. Konkret: Die – wie auch immer ermittelten – Summenwerte für die einzelnen Universitäten oder Fächer werden als Punktwerte interpretiert und behandelt – wie das ähnlich häufig im Umgang mit Testergebnisse in der Psychodiagnostik geschieht, wenn z. B. gesagt wird, ein Proband habe einen IQ

von 110 oder allgemeiner ein Testergebnis von x-Punkten. Messtheoretisch korrekt ist dagegen nur die Aussage zulässig, dass – wegen der möglichen Messfehler – wahre 'wahre' Werte mit einer bestimmten Wahrscheinlichkeit in einem Vertrauensbereich von + - x Punkten um den gemessenen Wert liegt. Da nicht gesagt werden kann, wo in diesem Bereich der 'wahre' Wert anzunehmen ist, bedeutet das konkret, dass Messungen, die zu unterschiedlichen Messwerten führen, nur dann als unterschiedlich gelten können, wenn sich die Vertrauensbereiche nicht überschneiden. Kritisiert wird entsprechend, dass die SPIEGEL-Studien Rangfolgen präsentieren, wo statistisch überhaupt keine Unterscheidungen möglich sind.

Status und Gewicht der Kritikpunkte:

Repräsentativität

Wie man zunächst einmal die notwendige Stichprobengröße berechnet, um mit einer festgesetzten Wahrscheinlichkeit von der untersuchten Stichprobe auf die jeweilige Grundgesamtheit schließen zu können, ist jedem Statistiklehrbuch zu entnehmen – und nicht Gegenstand individueller Einschätzungen. Spielräume und Diskussionsbedarf bestehen dort, wo entschieden werden muss, wieviel Unsicherheit man unter Abwägung aller Untersuchungsbedingungen (z. B. Kosten) für akzeptabel hält. Deutlich schwieriger ist demgegenüber die Stichprobenziehung bzw. die Beurteilung der Frage, welches die wesentlichen Merkmale sind, in denen die Stichprobe mit der Grundgesamtheit übereinstimmen muss, wie weit diese Übereinstimmung reichen muss und wie man sie feststellt. Neben klaren Kriterien, die auf vorangegangenen Untersuchungen basieren, kommen hier auch viele mehr oder weniger reflektierte Plausibilitätserwägungen und Routinen ins Spiel. Unstrittig ist aber, dass hierzu im Vorfeld der Erhebung Überlegungen und Direktiven notwendig sind und es nicht z. B. der Willkür einzelner Interviewer überlassen werden darf, welche Chancen (in diesem Fall) einzelne Studierende haben, Teil der Stichprobe zu werden.

Relationierung von Daten

Ob Daten sinnvoll zueinander in Beziehung zu setzen sind, ist Frage einer sinnstiftenden Theorie und insofern nicht in der Form klar zu entscheiden, wie z. B. im Falle der Repräsentativität die Größe der notwendigen Stichprobe. Die Entscheidung darüber, ob Daten sinnvoll relationiert werden können/sollten, ist aber forschungslogisch und -praktisch Stichprobenerwägungen vorgeordnet. Denn wo Relationen theoretisch sinnlos sind, erübrigt sich die Frage, wie oft sie wo untersucht werden sollen.

Ranking

Von den drei genannten Kritikpunkten ist dieser der methodologisch präziseste, gleichzeitig aber forschungspraktisch und -logisch nachgeordnet. Darüber, wie Befunde darzustellen und zu interpretieren sind, lohnt eine Diskussion erkennbar erst dann, wenn aussagekräftige Daten vorliegen. Dann aber stellt ein methodischer Fehler in diesem Punkt nicht die gesamte Untersuchung in Frage, sondern führt zu einer

Korrektur der Reichweite der Resultate: Man kann dann z.B. nicht scharf zwischen einzelnen Universitäten differenzieren, sondern nur zwischen Extremgruppen.

Andere mögliche Kritikpunkte, die sich z. B. auf die Formulierung von Fragen, Skalierung von Urteilen, Gewichtung von Urteilen usw. beziehen, sollen hier vernachlässigt werden, weil in diesen Punkten die methodologische Diskussion weit weniger zu klaren Kriterien und akzeptierten Prozeduren führt. Die oben angesprochenen Punkte haben demgegenüber den Vorteil, dass man von relativ klaren und unstrittigen Verfahrensregeln ausgehen kann, deren Beachtung sich kontrollieren lässt.

Insgesamt läuft bereits die oben skizzierte Kritik darauf hinaus, man habe Daten von zu wenigen, vermutlich noch dazu den falschen, Leuten erhoben, diese Daten in unsinnige Beziehungen gebracht und dann auch noch regelwidrig dargestellt und willkürlich interpretiert.

4. Reaktionen auf die Kritik an den SPIEGEL-Studien 1989/90 und 1992/93

Gemäß den gängigen Idealvorstellungen von empirischer Forschung, wie sie z. B. in der Auseinandersetzung zwischen quantitativer und qualitativer Forschung noch einmal expliziert und bekräftigt worden sind, wären nun – wenn denn die Kritik nicht entkräftet werden kann – bei solch klaren methodische Fehlern radikale Korrekturen des Forschungskonzeptes zu erwarten. Welche Konsequenzen die oben skizzierte Kritik tatsächlich hat, kann im vorliegenden Fall auf zweierlei Weise nachvollzogen werden: Einmal hat sich ein Mitarbeiter der ersten SPIEGEL-Studien, HORNBOSTEL (vgl. HORNBOSTEL/DANIEL 1995)[1], ausführlich zur methodologischen Kritik geäußert, zum anderen liegt mittlerweile eine weitere SPIEGEL-Studie (1998) vor (an der HORNBOSTEL wiederum beteiligt gewesen ist), an der geprüft werden kann, ob die Kritik an problematischen Verfahrensschritten zu Konsequenzen geführt hat.

Erwiderung

Bemerkenswert an der Auseinandersetzung HORNBOSTELS und DANIELS (1995) mit der Kritik an den SPIEGEL-Studien sind zunächst die Kriterien, die sie eingangs zur Beurteilung der Güte dieser Untersuchungen einführen:

- Das Lob einer Journalistin in der Wochenzeitung „DIE ZEIT".
- Der beträchtliche Aufwand, der mit der Befragung von mehr als 10000 Studenten getrieben wurde.
- Die „Auswirkungen der ersten SPIEGEL-Studie auf das Studienortswahlverhalten der Studienanfänger" (1995, S. 30).

Erkennbar orientieren sich die Autoren bei dieser positiven Beurteilung der Studien nicht an wissenschaftlichen Kriterien. Dem entspricht es, dass die Journalistin na-

[1] Für die Überlegungen im vorliegenden Text geht es allerdings nicht um eine persönliche Zurechnung der Studie(n).

mentlich genannt und zitiert, aber nicht einmal erwähnt wird, dass es begründete Kritik aus dem Kreis der Fachkollegen gegeben hat. Die Rede ist stattdessen von ‚pauschaler' Kritik, deren Berechtigung angesichts der o.g. klar erkennbaren Vorzüge der Studien man prüfen wolle. Kollegiale Kritik, die sich auf diese Studien bezieht, wird dann auch im weiteren Verlauf nicht explizit zur Kenntnis genommen, weder Namen, noch Veröffentlichungen, schon gar nicht konkrete Argumente. Der wenig argumentativen Einleitung, die sich zudem dezidiert an außerwissenschaftlichen Maßstäben orientiert, folgt dann allerdings doch eine Auseinandersetzung mit Problemen, die nicht näher genannte Kritiker angesprochen hätten.

Repräsentativität

Der Kritik an zu kleinen Stichproben (12 bzw. 18 Studierende pro Fach) wird entgegengehalten, dass die Grundgesamtheit, auf die man habe schließen wollen, nicht alle Studierenden umfasse, sondern nur die zwischen dem 5. und 10. Fachsemester, und das seien je nach Fach nur 20% bis 40% der Studierenden. Damit wird das Größenverhältnis der untersuchten Stichprobe zur Grundgesamtheit natürlich verbessert. Ob auch die Kritiker dies dann als den „Standards der Umfrageforschung" entsprechend (ebd., S. 31) akzeptieren würden, kann hier offen bleiben, festzuhalten ist immerhin, dass auch die Autoren im Kern die Kriterien akzeptieren, an denen die Kritik sich orientiert, und zu zeigen versuchen, warum im gegebenen Fall die Kritik nicht trifft.

Sie befassen sich ebenso mit dem o.g. zweiten Aspekt, nämlich der Frage der Stichprobenziehung. Sie räumen hier ein, dass es sich bei der Stichprobe nicht um eine Zufallsstichprobe gehandelt habe, also nicht jeder Studierende die gleiche Chance gehabt habe, Teil der Untersuchung zu werden. Die Wahl der Befragungsorte – Labor, Hörsaal, Bibliothek – bevorzugt erkennbar die Studierenden, die vor Ort die offiziellen Angebote der Universität nutzen (müssen), und klammert damit z.B. die Urteile der Studierenden systematisch aus, die mit diesen Angeboten nichts anzufangen wissen – etwa mit der Ausstattung der Bibliothek – oder die Möglichkeit haben, ihr Studium abseits von Hörsälen und Labors zu gestalten.

Es wäre (in anderen Kontexten) durchaus der Überlegung wert, ob für eine Beurteilung der Studienbedingungen in einzelnen Fächern die Urteile der hier selektierten Studierenden ausreichen. Dafür wären z. B. die, die sich vom Unibetrieb zurückziehen, wohl mindestens ebenso wichtig – aber natürlich weniger bequem zu erreichen. Hier reicht es allerdings festzustellen, dass HORNBOSTEL und DANIEL im Kern die Kriterien als verbindlich akzeptieren, von denen auch die Kritiker ausgehen. Sie bemühen sich zur Entkräftung der Kritik einerseits ihre Bemühungen um eine methodisch korrekte Durchführung der Stichprobenziehung zu belegen, andererseits Grenzen des Vorgehens zu diskutieren und in seinen Konsequenzen abzuschätzen. Ob ihre Argumente nun für alle Fachkollegen überzeugend sind, muss dabei nicht weiter interessieren, für den hier verfolgten Argumentationsstrang ist vielmehr wesentlich, dass sich Forscher und Kritiker den gleichen Kriterien verpflichten – wenn sie auch (wenig erstaunlich) Beurteilungsspielräume nutzen, für den jeweils eigenen Zugang zu werben.

Ranking

Der Versuch, als Resultat eine klare Rangreihe einzelner Universitäten zu präsentieren, wird von HORNBOSTEL und DANIEL zwar als aus journalistischer Sicht verständlich, statistisch aber nicht seriös dargestellt. Sie befürworten wegen der geringen Meßwertunterschiede benachbarter Universitäten/Fachbereiche – d.h. überlappende Vertrauensbereiche – eine Differenzierung klar unterscheidbarer Extremgruppen. Auch diese Beurteilung orientiert sich an gängigen methodologischen Kriterien.

Relationierung von Daten

Dieser Punkt steht jetzt am Ende der Darstellung, weil die Auseinandersetzung von HORNBOSTEL und DANIEL mit der methodologischen Kritik eine bemerkenswerte Lücke aufweist: Das forschungslogische vorgeordnete und deshalb nicht gerade unwichtige Problem, ob denn Urteile der Studierenden verschiedener Universitäten sinnvoll verglichen bzw. innerhalb einer Skalierung relationiert werden können, wird nicht behandelt. HORNBOSTEL und DANIEL behaupten in diesem Kontext nur, dass die Bedeutung derartiger Studentenbefragungen (in Abgrenzung etwa zu Professorenbefragungen) darin bestehe, dass „bei der Diagnose von Stärken und Schwächen eines Fachbereiches auf die Klientenperspektive in Form vergleichbarer Daten aus allen anderen Hochschulen im gleichen Fach zurückgegriffen werden kann" (ebd. S. 42).

Die SPIEGEL-Studie 1998: Uni-Test Europa

Diese positive Einschätzung macht dann auch verständlich, warum die Studie des SPIEGEL unter Mitarbeit von HORNBORSTEL aus dem Jahr 1998 diese Vergleiche fortsetzt und jetzt sogar auf Europa ausdehnt.

Neben dieser Ausdehnung des Vergleichs auf Europa wird als Verbesserung hervorgehoben, dass nun pro Fachbereich mehr Studierende als in den früheren Studien befragt wurden, nämlich ca. 50 (DER SPIEGEL 1998, S. 98). Da aber jetzt nicht mehr von der Beschränkung auf einen Teil der Studierenden (5.-10. Fachsemester) die Rede ist, dürfte sich allein durch diese Erhöhung der Zahl der Befragten gegenüber den früheren Studien tatsächlich am Verhältnis der Grundgesamtheit zur Stichprobe nichts ändern. Was die Sicherheit des Schlusses von der Stichprobe auf die Grundgesamtheit angeht, wird diese Veränderung für sich genommen also wohl kaum eine Verbesserung bringen.

Ebenfalls kein Fortschritt ist – auch im Anschluss an HORNBOSTEL und DANIEL (s.o.) – die Darstellung der Ergebnisse in einer Rangreihe, die den Eindruck vermittelt, man könne inhaltlich und rechnerisch einen klaren Unterschied zwischen einer Hochschule mit einem Gesamtpunktwert von 54 und einer mit 53 Punkten machen. Wichtiger als HORNBOSTELs methodologisch begründete Ablehnung einer solchen unseriösen Darstellung (s.o.) ist hier letztendlich wohl das „journalistisch reizvolle Siegertreppchen" (HORNBOSTEL/DANIEL 1995, S. 37) gewesen. Damit werden hier Aussagen über Unterschiede gemacht, über die man überhaupt nichts wissen kann.

Wichtiger als dieser Rückschritt in der Beurteilung und Darstellung der Resultate ist aber die Beibehaltung des Grundkonzepts: An verschiedenen Universitäten werden Studierenden ausgewählter Fachbereiche Fragen zu ihrer Studiensituation gestellt, und diese Beurteilungen werden dann miteinander verglichen. Zur Illustration am Beispiel der Europa-Studie: Da werden Ingenieurstudenten in Stockholm gebeten, auf einer Skala von 1 bis 6 das Bücherangebot zu beurteilen. Die Einzelurteile werden dann zum Durchschnittswert von 3,3 verrechnet. Ebenso werden Ingenieurstudenten in Darmstadt gefragt, der Durchschnittwert hier: 2,7, Studierende in Athen, der Durchschnittswert: 3,8 usw. Der wichtige Punkt ist nun, dass diese Urteile so behandelt werden, als habe der Studierende in Stockholm sein Urteil in Kenntnis der Bibliotheksausstattung in Athen, Darmstadt usw. abgegeben. Das aber kann ausgeschlossen werden. Die Studierenden werden auch zu einem derartigen Vergleich nicht aufgefordert.

Das resultierende Problem: Es ist ungewiss, ob sich das Urteil eines Studierenden in Athen über seine Bibliothek vor Ort überhaupt sinnvoll zu dem formal ähnlichen Urteil eines Studierenden in Darmstadt in Beziehung setzen lässt, und – wenn ja – mit welcher Skalierung. Das aber wird in den SPIEGEL-Studien immer wieder ohne Problematisierung unterstellt, und nur auf der Basis dieser Unterstellung kommt es dann zu Aussagen der Art, die TU Kopenhagen werde in den Ingenieurwissenschaften besser beurteilt als die Universität Padua. Wird aber diese Unterstellung nicht akzeptiert, erübrigen sich Diskussionen über Größe und Repräsentativität von Stichproben, Vertrauensintervalle usw., weil nur Zahlen mit unbekannter Bedeutung verarbeitet werden.

5. Zu möglichen Wirkungen auf die wissenschaftliche Forschung

WATZLAWICK (z. B. 1983, S. 27f) verwendet in anderen Zusammenhängen gern die Geschichte vom Mann, der nachts unter einer Laterne nach etwas sucht. Ein Passant will bei der Suche helfen und fragt den Mann, was er denn verloren habe, und ob er denn sicher sei, dass es genau hier geschehen sei. Der antwortet, er habe seinen Schlüssel verloren, und dann: „Nein, nicht hier, sondern dort hinten – aber dort ist es viel zu finster." Grotesk wird dies Verhalten, weil das Mittel, die Suche, vom Zweck abgekoppelt und unter der Hand zum Selbstzweck wird. Es geht nicht mehr darum, etwas zu finden, sondern darum, möglichst komfortabel zu suchen. Der entscheidende Mangel der Geschichte, so wie WATZLAWICK sie erzählt, besteht für mich darin, dass er einen Betrunkenen unter der Laterne suchen lässt – so als sei die Geschichte nur vorstellbar, wenn man eine verminderte geistige Funktionsfähigkeit des Suchenden unterstellt. Wesentlich interessanter wird die Geschichte, wenn man probeweise die Suche unter der Laterne als nicht eben seltenen Vorgang ansieht – der durchaus nicht als sonderbar auffallen muss. Die SPIEGEL-Studie scheint mir ein Beispiel für eine solche Suche unter der Laterne zu sein: Die Suche ist nämlich sinnlos, wenn eine Relationierung der erhobenen Daten sinnlos ist. Die Diskussion und Verbesserung

der Stichprobenziehung und anderer nachgeordneter Probleme entspricht dann allenfalls dem Bemühen, die Suchstrategien unter der Laterne zu verfeinern. Und die Europa-Studie verkündet, so betrachtet, dann vor allem, man habe nun noch sorgfältiger unter noch mehr Laternen gesucht.

Entsprechend ist bei gleichem Untersuchungsansatz auch eine Fortsetzung der Studien – etwa mit „Uni-Vergleich Europa-Amerika" oder gleich „Die beste Uni der Welt" – überraschungsfrei. Eine detaillierte methodische Analyse erübrigt sich ebenso wie eine Diskussion der Ergebnisse, weil der Untersuchungsansatz systematisch Daten mit unbekanntem Sinn produziert.

Wenn oben nach der Bedeutung methodologischer Kriterien in der Durchführung und Diskussion der SPIEGEL-Studien gefragt wurde, orientierte sich diese Frage zunächst an den Kriterien, die für wissenschaftliche empirische Forschungsvorhaben anzuwenden wären – und folgt damit dem Anspruch der Studien, wissenschaftlich seriöse Informationen bereitzustellen. Unter diesem Anspruch müssen die Studien als gänzlich verfehlt beurteilt werden. Denn unübersehbar werden wesentliche methodologische Standards dem Interesse an den gewünschten präsentablen Ergebnissen geopfert: Gewollt ist der (für Laien) möglichst einfach nachvollziehbare **Vergleich** von Universitäten. Gegenüber obskuren Indizes, die Bücherbestände, Drittmittel, Studenten und Mensapreise verrechnen, werden mutmaßlich überzeugendere Betroffenen-/Expertenurteile als Basis des Vergleichs herangezogen. Diese Daten sind nun zwar eingängiger, aber für den beabsichtigten Vergleich ungeeignet – Untersuchungszweck und Untersuchungsverfahren/-daten passen nicht zusammen. Zur Lösung des Problems könnte man nun entweder den Zweck (Univergleich) beibehalten und die Verfahren anpassen (vergleichbare Daten erheben) oder aber das Verfahren der Studentenbefragung beibehalten und auf den Vergleich der Universitäten verzichten. Untersuchungszweck könnte dann z. B. eine differenzierte Rückmeldung für einzelne Fachbereiche sein – allerdings wären auch dafür die erhobenen Daten ungeeignet, weil (mit Blick auf den Vergleich) zu pauschal. Mit Blick auf die angestrebte öffentliche Wirkung sind vermutlich beide Lösungen unbefriedigend. Jedenfalls besteht die ‚Lösung' hier darin, an der ursprünglichen Zielsetzung festzuhalten und das methodologische Problem einfach nicht zur Kenntnis zu nehmen (vgl. a. SPIES 1989; DRERUP 1989).

Vom Auftraggeber der Studie aus betrachtet ist das durchaus konsequent, da die Wirkungen, die für den SPIEGEL vor allem interessant sein dürften, nicht davon abhängig sind, ob die Daten der Untersuchungen nach methodologischen Standards einen Sinn ergeben können. Dafür reicht vielmehr bereits der Anschein von Sinnhaftigkeit und Relevanz, also Face-Validity oder „psychologische Validität" (LIENERT 1969, S. 47), vollkommen aus. Das Image des wachen Medien-Gewissens auch und gerade dort, wo andere gesellschaftliche Instanzen versagen, wird für die Öffentlichkeit sicherlich eher durch den betriebenen Aufwand (viele Befragte, Universitäten, Länder) und die Beteiligung von Fachleuten als durch methodologische Sauberkeit gepflegt. Der Zweck wird also bereits durch die publikumswirksame Suche unter der Laterne erreicht – das Finden ist nicht mehr nötig.

Ebenso wird man davon ausgehen können, dass der erwünschte Druck auf die Bildungspolitik und die Universitäten, Ressourcen kontrollierter und reflektierter einzusetzen, nicht von der methodologischen Güte der Untersuchungen abhängen wird (vgl. SPIES 1989, DRERUP 1989). Auch so schaffen Studien dieser Art einen (zunächst) diffusen Rechtfertigungsdruck, der möglicherweise gerade dadurch effizient ist, dass er diffus und damit potentiell willkürlich in der Auslegung und Anwendung bleibt und im Zweifel die Legitimation für Sanktionen, die Umverteilung und Kürzung von Mitteln, Streichung von Studiengängen usw. schafft.

Man könnte nun dem von KROMREY referierten Vorschlag folgen und Evaluation als grundsätzlich von wissenschaftlicher Arbeit verschieden anzusehen: „Während in wissenschaftlich angelegten Vorhaben methodologische Standards von ausschlaggebender Bedeutung seien, stehe für Evaluationsvorhaben das Interesse an nützlichen Informationen im Blickpunkt" (1995a, S. 333). Zu ergänzen ist wohl: Und diesen Interesse würden dann eben bei Bedarf methodologische Standards geopfert. So deutlich sich dies Interesse und die resultierenden Konsequenzen im hier diskutierten Fall zeigen lassen, so wenig sinnvoll scheint mir die vorgeschlagene Trennung zwischen methodologisch fragwürdiger Auftragsforschung unter dem Diktat der Nützlichkeit und der wissenschaftlichen Forschung, die sich methodologischen Standards verpflichtet – im Übrigen folgt auch KROMREY dieser Empfehlung so nicht.

Im Folgenden konzentriere ich mich auf zwei Schwierigkeiten, diese Trennung vorzunehmen:

- Erstens hat empirische Forschung im Auftrag der Medien – wiederum unabhängig von ihrer methodologischen Güte – Aus-/Rückwirkungen auf wissenschaftliche Forschung. Sie produziert etwa ein Bild/Bilder wissenschaftlicher Forschung und etabliert Erwartungshaltungen hinsichtlich der Leistungsmöglichkeiten von Forschung – in welcher Zeit welche Fragen wie beantwortet werden können. Die Frage danach, in welcher Weise wissenschaftliches Denken popularisiert werden kann und z. B. Einfluss auf das Alltagsdenken und -handeln gewinnt, ist also sinnvoll durch die Frage nach Einflüssen in der umgekehrten Richtung zu ergänzen, danach, wie die Wissenschaftspraxis von öffentlichen Erwartungen beeinflusst wird.
- Zweitens sind durchaus begründete Zweifel daran angebracht, dass wissenschaftliche Forschung als methodologischen Standards folgend klar z. B. von Auftragsforschung unterschieden werden kann. Hier ist etwa im Kontrast zu den SPIEGEL-Studien wesentlicher differenzierter zu überlegen, was Forschung als wissenschaftlich ausweist.

Zum ersten Punkt: Zunächst ist – wie auch in anderen Fällen gehäufter Regelverstöße – mit Gewöhnungseffekten zu rechnen. Zur Prüfung, wie wirksam die im gegebenen Fall schon sind: Man stelle sich eine Untersuchung vor, die belegen will, wo es „Das beste Frühstück Europas" gibt, und dazu 7000 Europäer in 15 Ländern jeweils ihr Frühstück auf sechsstufigen Ratingskalen nach Geschmack, Sättigung, Preis, Nährstoffen, Aussehen usw. beurteilen lässt. Während diese Studie vor allem Heiterkeit erzeugen dürfte, sind im Ansatz gleichartige Studien zur Evaluation der Hochschulen

inzwischen schon üblich, und man darf wohl davon ausgehen, dass sie demnächst dann mit Blick und Verweis auf die Vorläuferstudien als ‚bewährt' gelten (vgl. z. B. a. FEYERABEND 1986, S. 23). Dass derart abnehmende Sensibilität gegenüber Regelverstößen und die Gewöhnung an unsinnige Abläufe und Verfahren nicht auf die Laienöffentlichkeit beschränkt bleibt, ließe sich auch für den Bereich empirischer wissenschaftlicher Forschung leicht zeigen (s. u.).

Gerade wegen ihrer Betonung wissenschaftlicher Seriosität dürften derartige Studien das Bild der Öffentlichkeit von den Leistungsmöglichkeiten wissenschaftlicher Forschung beeinflussen. Danach erscheinen dann in diesem Fall klare und eindeutige Unterscheidungen der Universitäten möglich. Das ist insbesondere deshalb wichtig, weil dies vollkommen unrealistische Bild dann die Erwartungen an weitere Studien prägen und – spätestens über die Vergabe von Forschungsmitteln – einen entsprechenden Druck auf die Forscher ausüben dürfte, wiederum in ‚bewährter' Manier erfolgreich zu sein und ‚fast research' zu betreiben.

KRIZ (1995) beurteilt die Gefahr vollkommen unrealistischer Erwartungen an die Leistungsmöglichkeiten wissenschaftlicher Forschung allerdings als relativ gering, weil durch die Veröffentlichung unterschiedlicher konkurrierender Untersuchungen mit unterschiedlichen Designs und unterschiedlichen Ergebnissen „der Eindruck einer 'wahren, wirklichen' Rangreihe auch für Laien fragwürdig und das Bedürfnis nach einer platten Antwort auf die Frage: ‚Welche Uni ist die beste?' enttäuscht" werde (S. 16). Erstens wäre dies aber keine Enttäuschung, die mit einem Verständnis der Grenzen wissenschaftlicher Arbeit einhergeht, sondern wohl eher den Eindruck vermittelt, man müsse sich aus den angebotenen eben die passende Wahrheit aussuchen. Zweitens bedeutet eine Enttäuschung der Hoffnung auf Eindeutigkeit und Klarheit – vor allem wenn die Gründe nicht verstehbar sind – nicht, dass damit die Erwartung, wissenschaftliche Untersuchungen könnten letzte Wahrheiten aufdecken, aufgegeben wird.

Erwartbar und problematisch scheint mir insbesondere, dass Gewissheiten, die nach Art der SPIEGEL-Studien hergestellt werden, durch methodologisch seriöse(re) Untersuchungen nicht zu korrigieren sind, da diese entweder mehr Fragen als Antworten liefern (vgl. z. B. KROMREY 1995b) oder ihre Befunde mit allerlei Einschränkungen versehen, hinsichtlich ihrer psychologischen Validität für die Öffentlichkeit jedenfalls nicht mit Studien nach SPIEGEL-Art konkurrieren können.

So sehr durch Medienstudien der hier diskutierten Art auf der einen Seite systematisch vereinfachende Vorstellungen, falsche Erwartungen und damit Enttäuschungen produziert werden, so sehr wird aber auf der anderen Seite die Notwendigkeit empirischer Forschung unterstrichen. So ebnen sie – gerade weil die öffentlichkeitswirksamen Großstudien letztlich keine konkreten Fragen beantworten – auch den Weg für kleinere, weniger spektakuläre Studien vor Ort, die dann z. B. Studienverläufe, Lernerfolge in Veranstaltungen, Prüfungserfolge usw. evaluieren.

Zum zweiten Punkt: Die Missachtung methodologischer Standards als besonderes Merkmal von Auftragsforschung/Evaluationsstudien einzustufen, schon gar wissenschaftliche Forschung davon grundsätzlich zu unterscheiden, ist deshalb wenig sinn-

voll, weil sich auch im Bereich wissenschaftlicher Forschung an zahlreichen Beispielen (vgl. z. B. KRIZ 1981; RITTELMEYER 1992) nicht nur fahrlässige, sondern routinemäßige Verstöße gegen weitestgehend akzeptierte methodologische Normen belegen lassen. Da es sich dabei nicht nur um Regelverstöße in Detailfragen handelt, sondern durchaus auch um solche, die sinnvoll interpretierbare Befunde zweifelhaft werden lassen – etwa der kreative Umgang mit Skalenniveaus, die Weiterverwendung indiskutabler (Fragebogen-)Items, die Reifizierung von Faktoren – wäre eher zu vermuten, dass auch in der wissenschaftlichen Forschung das Suchen unter Laternen zu den durchaus respektablen Tätigkeiten gehören kann – wobei je nach ‚Gemeinde'zugehörigkeit jeweils unterschiedliche Methodenfehler de facto akzeptiert und von der Kritik ausgenommen werden.

Auch hier darf man ein Interesse an nützlichen Ergebnissen annehmen, das im Zweifel methodologischen Bedenken vorgeht – nur eben ein anderes als z.B. in den SPIEGEL-Studien. Allerdings müssen dabei ‚nützliche Ergebnisse' nicht darin bestehen, konkrete Erkenntnisse über empirische Gegebenheiten zu produzieren, sondern können vielleicht schon dann erzielt werden, wenn man der Fachöffentlichkeit demonstriert, dass man bestimmte Forschungs-/Suchmethoden gekonnt anwenden kann. Wo z.B. Untersuchungen routinemäßig schneller und zahlreicher abverlangt werden, als sie unter Beachtung methodologischen Normen produziert werden können, nehmen ‚Vereinfachungen' zu. Etwa, indem man – wie das HORNBOSTEL und DANIEL tun – methodologische Probleme, mit denen in der Untersuchung nicht angemessen umgegangen wird, vollständig unerwähnt lässt und stattdessen differenziert die Probleme diskutiert, zu denen man etwas zu sagen hat, oder indem man diese Probleme zwar anspricht, aber erst nach der Untersuchung in der Diskussion der Ergebnisse – so, als stelle sich das Problem erst jetzt. Diese Variante verbindet dann eine nicht unnötig komplizierte Untersuchung mit der (nachträglichen) Demonstration methodologischen Problembewusstseins.

Kurz: Auch in der wissenschaftlichen Forschung sollte man davon ausgehen, dass methodologische Standards beachtet werden, soweit sie von der jeweiligen Referenzgruppe als wichtig angesehen werden und solange dadurch nicht der notwendige Output behindert wird (vgl. dazu allgemeiner: FEYERABEND 1986).

Der Unterschied zu Evaluationsstudien der hier diskutierten Art wäre danach kein grundsätzlicher, sondern eher ein gradueller. Insbesondere wo mediengerechte Aufbereitung und Öffentlichkeitswirkung Bedeutung gewinnen, mögen zunehmend methodologische Standards in Vergessenheit geraten (vgl. a. RITTELMEYER 1992) – bis hin zur Produktion methodologisch vollkommen unsinniger Daten passend zur jeweiligen Überschrift.

Zur Charakterisierung wissenschaftlicher Forschungspraxis reicht es dann aber nicht aus, allgemein die Beachtung methodologischer Standards zu beschwören. Gesagt werden müsste vielmehr genauer, gegen welche methodologischen Standards man wann durchaus verstoßen darf, ohne als inkompetenter Forscher zu gelten. Welche Fehler man vielleicht sogar begehen muss, um als kompetenter Forscher akzeptiert zu werden. Untersuchungen wie die SPIEGEL-Studien sind in dieser Hinsicht auf eine fruchtbare Weise unbequem, indem sie einerseits (wieder einmal) die Schwierigkeiten deutlich machen, in der Wissenschaftspraxis tatsächlich gültige methodologische Standards anzugeben, und andererseits die Notwendigkeit deutlich machen, unterhalb von praxisfernen Lehrbuchpostulaten konkreter zu fassen, wo wissenschaftliche Forschung aufhört.

Die möglichen Auswirkungen der hier skizzierten Wissenschaftsforschung im Auftrag der Medien auf die Praxis empirischer Forschung ergeben nach den hier vorgestellten Überlegungen ein uneinheitliches Bild. Einmal aktualisiert und verschärft sie das innerwissenschaftliche Unbehagen mit dem eigenen Selbstverständnis, indem sie öffentlichkeitswirksam ein Talmibild von Wissenschaft nicht nur präsentiert, sondern darüber hinaus zur Orientierung nahelegt. Sie erzeugt zudem einen zusätzlichen Druck, im Zweifel ökonomische Durchführung und wirksame Präsentation höher als methodologische Seriosität zu gewichten. Während so in der Tendenz ein eher nachlässiger Umgang mit methodologischen Standards gefördert wird, stimuliert auf der anderen Seite die mediale Präsentation und Aufwertung empirischer Forschung die Nachfrage nach empirischen Untersuchungen und Befunden, die zunehmend unverzichtbarer Bestandteil der Legitimation von Entscheidungen werden – auch wenn man ihnen immer weniger vertraut und sie immer schneller durch aktuell passende ersetzt.

Literatur

DER SPIEGEL: Studieren heute. Welche Uni ist die beste? In: SPIEGEL SPECIAL (1990), H.1.

DER SPIEGEL: Welche Uni ist die beste? In: DER SPIEGEL (1993), H.16, S. 80-101.

DER SPIEGEL: Top in Europa. Die besten Hochschulen. In: SPIEGEL SPECIAL (1998), H.6, S. 77-81.

DER SPIEGEL: Uni-Test Europa. In: DER SPIEGEL (1998), Teil 1: H.19, S. 94-117 , Teil 2: H.20, S. 72-91.

DRERUP, H.: Probleme außerwissenschaftlicher Verwendung von Erziehungswissenschaft. Zum Einfluß von Erziehungswissenschaft im politisch-administrativen Bereich. In: KÖNIG, E./ZEDLER, P. (Hrsg.): Rezeption und Verwendung erziehungswissenschaftlichen Wissens in pädagogischen Handlungs- und Entscheidungsfeldern. Weinheim 1989, S. 143-166.

FEYERABEND, P.: Wider den Methodenzwang. Frankfurt a.M. 1986.

HORNBORSTEL, S./DANIEL, H.-D.: Das SPIEGEL-Ranking. Mediensensation oder ein Beitrag zur hochschulvergleichenden Lehrevaluation? In: MOHLER, P. P. (Hrsg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung. 2. überarb. Aufl. Münster/New York 1995, S. 29-44.

KRIZ, J.: Methodenkritik empirischer Sozialforschung. Eine Problemanalyse sozialwissenschaftlicher Forschungspraxis. Stuttgart 1981.

KRIZ, J.: Die Wirklichkeit von (Vor-)Urteilen). Über die inhaltlichen und methodischen Hintergründe der STERN-Image-Analyse. In: MOHLER, P. P. (Hrsg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung. 2. überarb. Aufl. 1995 Münster/New York, S. 11-28.

KROMREY, H.: Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie 15(1995), H.4, S. 313-336. (a)

KROMREY, H.: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: MOHLER, P. P. (Hrsg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung. 2. überarb. Aufl. Münster/New York 1995, S. 105-128. (b)

LIENERT, G. A.: Testaufbau und Testanalyse (durch einen Anhang über Faktorenalyse ergänzte Auflage). 3. Aufl. Weinheim/Berlin/Basel 1969.

MOHLER, P. P. (Hrsg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung. 2. überarb. Aufl. Münster/New York 1995.

RITTELMEYER, C.: Über die Fehldeutung von Forschungsdaten. Ein Kapitel über wissenschaftliche Verantwortung. In: Bildung und Erziehung 45(1992), H. 3, S. 355-366.

SPIES, W. E.: Aufnahme pädagogischen Wissens in der Kultusverwaltung. In: KÖNIG, E./ZEDLER, P. (Hrsg.): Rezeption und Verwendung erziehungswissenschaftlichen Wissens in pädagogischen Handlungs- und Entscheidungsfeldern. Weinheim 1989, S. 99-110.

WATZLAWICK, P.: Anleitung zum Unglücklichsein. München 1983.