DISKUSSION:

Von den Problemen anwendungsorientierter Sozialforschung und den Gefahren methodischer Halbbildung[1]

I.

Der vorliegende Text handelt von zwei Teilthemen, mit denen sich jeder konfrontiert sieht, der empirische Sozialforschung nicht lediglich als eine (bei den Studierenden meist unbeliebte) Hochschuldisziplin sieht, sondern der als Sozialwissenschaftler in der Praxis und für die Praxis empirische Informationen zu beschaffen und/oder zu verwerten hat.

Zum ersten Teilthema:
Von den Problemen anwendungsorientierter Sozialforschung.

Die gibt es in Fülle. Jeder anwendungsorientierte Forscher kennt die Ansprüche, denen sich Sozialwissenschaftler in der Praxis gegenübersehen:

- *Die Forschung soll*: wenig kosten, schnell gehen, wenig Aufwand erfordern; sie soll manchmal sogar nach erster fachmännischer Anleitung vom Laien selbst durchführbar oder zumindest weiterführbar sein.
- *Die Ergebnisse sollen* unmittelbar anwendungsrelevant sein, dazu möglichst einfach und möglichst allgemein verständlich, aber dennoch selbstverständlich nicht durch zu große Vereinfachung verzerrt.

Diesen Ansprüchen steht hinderlich entgegen, daß *soziale Sachverhalte immer komplex* sind, häufig nur dem Fachmann verständlich, schwer durchschaubar, selten direkt beeinflußbar.

Was kann bei dieser Konstellation passieren? Es können - vielleicht sogar: *es müssen* - sich Fehler einschleichen:

- *Fehler bei der Informationssammlung* durch Orientierung an der Forderung nach Zeit- und Aufwandminimierung (schnell, einfach, laienverträglich),
- *Fehler bei der Informationsaufbereitung* (möglichst 'einfache Statistik', Reduktion der Informationsfülle auf wenige 'Kennwerte', Verzicht auf komplexe Modelle),

---

[1] Der Beitrag basiert auf einem Vortrag vor der Fakultät für Sozialwissenschaft der Ruhr-Universität Bochum. Er folgt weniger dem Stil einer wissenschaftlichen Abhandlung als dem einer argumentativen Positionsbestimmung. Da zudem einer „Diskussion" eher der Diskurstyp der Rede entspricht, habe ich mich bemüht, den Redetext bei der Überarbeitung nicht unversehens in eine „Schreibe" zu transformieren.

- *Fehler bei den Schlußfolgerungen und Empfehlungen* (Orientierung an leichter Durchsetzbarkeit, an *einfachen* Handlungsempfehlungen, an der Sichtbarkeit des Eingreifens - Motto: Es muß etwas geschehen!).

Anwendungsorientierte Sozialforschung - obwohl sie häufig gegenüber der theorietestenden oder der Grundlagenforschung als das wissenschaftliche Aschenputtel betrachtet wird - *ist also keine einfache Sache.*

Zum zweiten Teilthema:
Von den Gefahren methodischer Halbbildung.

Wer an der Hochschule die meist undankbare Aufgabe hat, Methoden empirischer Forschung zu vermitteln, der kennt die typische Erwartung: Wasch' mir den Pelz, aber mach' mich nicht naß! Gewünscht wird:

- so wenig wie möglich davon, nur das "wirklich Notwendige", insbesondere:
- einfache Verfahren der Informationserhebung,
- einfache Verfahren der Auswertung, nur ganz 'einfache Statistik' (wenn überhaupt),
- nur das, was 'für die Praxis' wirklich erforderlich ist. (Alles andere ist allenfalls etwas für 'Methoden-Freaks')

Dies steht natürlich im Widerspruch zu der gerade skizzierten Gefahr, an vielen Stellen viele Fehler zu begehen. Es beschwört auch eine *ganz besondere Gefahr* herauf - nämlich daß die Gefahren methodischer Halbbildung *als solche* überhaupt nicht erkannt werden: *Je schlichter die methodische Ausbildung, um so überzeugter ist man von der Richtigkeit* (wenn nicht gar: 'Objektivität') *seiner Resultate.* Was alles schiefgehen kann, hat man nie kennengelernt; also besteht auch kein Anlaß für Zweifel an der schönen Gleichung *'einfach = gut'.*

II.
Ich habe zu Beginn *drei Bereiche* genannt, *in denen typischerweise gravierende Fehler unterlaufen können* (bzw. müssen):

- bei der Informationssammlung,
- bei der Informationsaufbereitung bzw. -auswertung und
- bei den Schlußfolgerungen und Empfehlungen.

Im folgenden sei dies an einigen *Beispielen* aus einem Bereich illustriert, der seit einigen Jahren in der Öffentlichkeit, in der Politik und auch an den Hochschulen in der Diskussion steht: der *"Qualität der Lehre" sowie ihrer Evaluation*. Spätestens seit einem Jahrzehnt sind auch die Bundes- und Landesministerien auf die Sozialwissenschaftler als anwendungsorientierte Forscher aufmerksam geworden und erwarten von ihnen nützliche Beiträge. Sie sollen einfache, in der Durchführung 'ökonomische', auch von Nichtsoziologen leicht anwendbare Modelle der empirischen Evaluation der Lehre anbieten.

- So heißt es zum Beispiel im *Aktionsprogramm Qualität der Lehre* der nordrhein-westfälischen Landesregierung 1991: 'Im Mittelpunkt einer Kritik des Ausbildungsangebots steht die Beurteilung der Lehrveranstaltungen und damit auch der Lehrenden (Veranstaltungskritik). Daneben kommt die Beurteilung des Prüfungsbetriebs und - mit gewissen Einschränkungen - der Studiengänge in Betracht.' (MWF NW 1991, 124)
- Und 1992 forderte der seinerzeitige Bundesbildungsminister Ortleb ein 'Anknüpfen von "Belohnungsmechanismen" an gute Leistungen in der Lehre oder die systematische Evaluation von Lehrleistungen der Hochschulen' - so eine seiner *'Thesen zur Belebung der Leistungskraft der Hochschulen'* (Informationen Bildung Wissenschaft, Nr. 7-8/92, S. 90-92).
- Ein weiterer Diskussionsstrang läuft an den Hochschulen selbst darauf hinaus, die knapp gewordenen (und auch künftig knapper werdenden) Finanzmittel „leistungsbezogen" zuzuteilen. So wünscht sich etwa an der Freien Universität Berlin das Präsidialamt ein an Indikatoren anknüpfendes, formelgebundenes Instrument, das es erlaubt, einen zunehmenden Anteil der noch verbleibenden *Haushaltsmittel nach Leistungskriterien* zu verteilen.

Gefragt sind also empirische und damit - so hofft man - quasi objektive Bewertungen nicht nur der Lehre, sondern des gesamten Ausbildungs- und Prüfungsbetriebs.

III.

Selbstverständlich verfügt die anwendungsorientierte Sozialforschung über ein geeignetes Design zur empirischen Bewertung (auch von Lehre und Ausbildung), nämlich das Design der *Programm-Evaluation*. Es ist allerdings methodisch aufwendig, verlangt vom Durchführenden Forschungserfahrung, kostet viel Zeit, setzt u.a. relativ genaue Kenntnisse der Ziele, Instrumente und Zielgruppen des Programms sowie den Konsens über geeignete Beurteilungskriterien voraus, muß selbstverständlich auch auf die Inhalte des zu bewertenden Programms zugeschnitten sein. Es ist also *alles andere als ein einfaches Verfahren*.

An ein solches differenziertes Forschungsprogramm ist denn auch gar nicht gedacht, wenn im geschilderten Kontext der Ruf nach Evaluation der Qualität von Lehre und Ausbildung laut wird. Gedacht wird nicht an gegenstandsspezifische, sondern an möglichst *generelle Kriterien für 'Qualität'*, an Kriterien, die gerade stoff- und fach- und lernzielunabhängig sind, die möglichst flächendeckend zum gleichen Zeitpunkt in der gesamten Universität anwendbar sein sollen.

Sagt dazu *der verantwortungsbewußte und methodisch reflektiert verfahrende Sozialforscher*, diese Aufgabe sei unlösbar, so scheint demgegenüber für den *methodisch Halbgebildeten* die Lösung ganz einfach: Man befrage - so deren Rat - die Studierenden in den jeweiligen Lehrveranstaltungen bzw. Organisationseinheiten und erhebe deren Bewertungen; und man tue dies anhand kurzer Fragebögen mit relativ globalen (also: inhaltsunabhängigen) Beurteilungen. Die Studierenden - dieses Argument liegt auf der Hand - sind als die von der Lehre und Ausbildung ganz konkret 'Betroffenen' am besten in der

Lage, aus eigener Erfahrung auch die Qualität des ihnen Gebotenen zuverlässig zu beurteilen.

Zwar ist ein solches Vorgehen - methodisch gesehen - kein Evaluationsdesign, sondern 'ganz normale' Umfrageforschung; dies wäre jedoch kein Problem, wenn die folgenden, in den Materialien eines hochschuldidaktischen Zentrums zu lesenden Ausführungen zuträfen:

- *Urteile (Schätzungen) von Studenten über die Lehre sind verläßlicher als Urteile der Dozenten über die Leistung der Studenten und ebenso verläßlich wie Urteile von Kollegen über die Lehre,*
- *sie legen praktisch dieselben Kriterien für gute Lehre an wie die Dozenten selbst,*
- *sie sind - wenn man etwa 20-30 Studenten urteilen läßt - zuverlässig wie professionelle Testverfahren,*
- *sie sind von anderen Merkmalen der Studenten selbst und der Dozenten wenig beeinflußt.* (Schmidt 1980, 51-52)

Obwohl bereits rund 20 Jahre alt, werden diese Argumente auch in der gegenwärtigen Diskussion um Lehrevaluationen immer noch vertreten. *Leider jedoch ist jede einzelne dieser Aussagen falsch*, was allerdings nicht ohne weiteres erkennbar ist - insbesondere dann nicht, wenn man auf die gewünschten 'einfachen Verfahren' der Datenerhebung und -auswertung zurückgreift. Zwar lernt man Studierende/r manches von den Gefahren und Fallstricken der empirischen Sozialforschung (insbesondere des Instruments Befragung) schon in den ersten Semestern, vergißt es dann aber erfolgreich wieder. Es ist erstaunlich, wieviel von ihrem *theoretischen* Wissen selbst gestandene Soziologen vergessen können, sobald es um die *praktische* Anwendung geht.

So gibt es denn auch ganz *aktuelle Beispiele* dafür, daß Sozialwissenschaftler Befragungsinstrumente vorlegen, die das im obigen Zitat erhoffte Wunder vollbringen sollen. Etwa ein von Psychologen entwickeltes und 1994 veröffentlichtes '*Heidelberger Inventar zur Lehrveranstaltungs-Evaluation*' (abgekürzt: HILVE) oder ein im 'Projekt Pro Lehre' der Freien Universität Berlin konzipiertes und wiederholt eingesetztes '*Studienbarometer*', das zu vergleichenden Aussagen über die Ausbildungsqualität an Fachbereichen und Instituten beitragen soll und dabei mit einer einzigen Seite auskommt.

'*Praxisorientiert*' sollen solche Verfahren *in mehrfacher Weise* sein: Sie sollen den Lehrenden Hinweise zur Optimierung ihrer Arbeit liefern; sie sollen die Abgrenzung der 'guten' von den 'schlechten' Lehrenden bzw. Fächern oder Fachbereichen ermöglichen (bis hin zum Ranking); sie sollen zur 'Objektivierung' der Zuteilung von Ressourcen auf der Basis von Qualitäts-Daten beitragen (Stichwort: Leistungsanreize). Damit ist ein weites Feld angesprochen sowohl hinsichtlich der Schwierigkeiten praxisorientierter Forschung als auch hinsichtlich der Gefahren, wenn dies mit methodischer Halbbildung angegangen wird (auch - und gerade! - wenn dabei nur 'ganz einfache' Verfahren eingesetzt werden).

IV.

Ich komme damit zum angekündigten *Beispiel für die erste Fehlerquelle* - **'Fehler bei der Informationssammlung' -** und illustriere dies am 'Studienbarometer' der FU Berlin.

Hier einfügen: Fragebogen FU-Studienbarometer

Als *Zweck des Studienbarometers* wird genannt: "... schnell und systematisch die *Einschätzung von Studierenden zu wichtigen Aspekten des Studienbetriebs* am Fachbereich" zu erheben. "Mit dem Studienbarometer werden Werte ermittelt, die einen Rückschluß auf die Einschätzung der Studierenden ermöglichen. Darüber hinaus können die Werte mit denen anderer Fachbereiche verglichen werden, und beim mehrmaligen Einsatz des Instruments läßt sich eine zeitliche Entwicklung erkennen" (PPL-Beiblatt, Sommersemester 1994).
Eine weitere Funktion: Die Daten sollen eine der *Informationsgrundlagen für Berichte der Fachbereiche über "Lehre und Studium und die Umsetzung der Studienstrukturreform"* sein: "In diesen Berichten sollen jeweils differenzierte Beurteilungen der Lehr-, Studien- und Prüfungsverhältnisse enthalten sein. Diese sollten ... dann zum Ausgangspunkt von konkreten Maßnahmen zur Verbesserung der Verhältnisse" gemacht werden (PPL-Schreiben vom 5. Oktober 1995).

Genau dieses aber ist von der Konzeption des Instruments und von der Erhebungssituation her *in keiner Weise gewährleistet.* Die Ergebnisse stellen *nicht die gewünschte differenzierte Einschätzung* der Lehr- und Studiensituation *durch* die Studierenden bereit, sondern sie liefern lediglich globale (also undifferenzierte) Stimmungsbilder *bei* den Studierenden.

Einige *Beispiele* mögen veranschaulichen, daß es mit einem Instrument wie dem "Studienbarometer" nicht möglich ist, vergleichbare *Bewertungen von Sachverhalten* zu erheben:

- ***'1. die didaktische Kompetenz der Lehrenden'*:**

  In keinem Fachbereich existieren '*die* Lehrenden' als Menge gleichartig lehrender Personen. Es existiert darüber hinaus nicht '*die* Didaktik' als Regelsystem, das für alle Studierenden und für jeden Lernstil in gleicher Weise angemessen wäre. Wer als Befragter diesen Sachverhalt bewerten soll, muß dies

  a) aus der eigenen Vorstellung (was Didaktik sei) und aus der eigenen subjektiven Perspektive tun, kann

  b) das Urteil nur über diejenigen Personen abgeben, die er als Lehrende selbst kennt und/oder über die er von anderen etwas gehört hat, muß

  c) aus diesen unterschiedlichen und teils widersprüchlichen Fragmenten einen 'subjektiven Mittelwert' bilden (für den keinerlei Konstruktionsvorschriften existieren, dessen Zustandekommen also von der Tagesstimmung abhängt).

- ***'2. die Vollständigkeit des Lehrangebots'*:**

  'Vollständigkeit' kann unter sehr unterschiedlichen Kriterien eingeschätzt werden (Fachsystematik, eigener thematischer Schwerpunkt; je Semester, verteilt über mehrere Semester; aktueller Stand des Fachs, historische Perspektive; etc.). Die Beurteilung setzt voraus, daß die bewertende Person einen Vergleichsmaßstab

besitzt, also weiß, was alles dazugehört (diesen Vergleichsmaßstab können Studierende aber noch nicht besitzen). Sie setzt ferner einen Überblick über das Gesamtangebot - also vollständige Information zum Zeitpunkt des Beantwortens der Frage - voraus (im Bewußtsein präsent sein kann aber nur ein zufälliger, sehr subjektiv gefärbter Ausschnitt aus dem Gesamtangebot).

- ***'3. die Übersichtlichkeit des Lehrangebots'*:**

  Je 'vollständiger' das Lehrangebot (je größer also die Auswahl*möglichkeiten*) und je *weniger* 'verschult' der Studiengang (je größer also die *Notwendigkeit* des Auswählens), desto 'unübersichtlicher' erscheint bei gegebenem Informationsverhalten das Angebot. Zugleich gilt: Je ausgeprägter das Informationsverhalten, desto 'übersichtlicher' erscheint ein vorhandenes Angebot. Was aber bedeutet dann ein im Fragebogen eingetragenes '+': Weniger Auswahl, daher übersichtlich? Verschulter Studiengang, daher keine Orientierungsprobleme? Hinreichendes Informieren der Studierenden, daher übersichtlich trotz Angebotsvielfalt? ...

Dies sind - willkürlich herausgegriffen - die ersten drei Items des Fragebogens. Ganz ähnliche Probleme tauchen bei jedem weiteren Item auf, *sofern die erhobenen Antworten als Bewertungen der angesprochenen Sachverhalte gedeutet werden sollen. In dieser Hinsicht* wären die Daten *Artefakte der* durch den Fragebogen geschaffenen *Erhebungssituation*.

Die genannten Kritikpunkte treffen dagegen *nicht* bei einer anderen Interpretation der Antworten zu, nämlich bei einer Interpretation als Aussagen über die *subjektive Zufriedenheit bzw. Unzufriedenheit* der Befragten. Allerdings geben Zufriedenheiten nicht Auskunft über die Qualität von Sachverhalten, sondern über die subjektive Stimmungslage der 'Betroffenen', über ihre individuelle Beziehung zu der angesprochenen Realität, *wie sie sie wahrnehmen*. Sie sind keine „Qualitätsmessung", wohl aber - und diese Assoziation vermittelt ja auch die Bezeichnung des Instruments - Indikator für vorherrschende Stimmungen (ähnlich dem aus den Medien bekannten „Politbarometer" der Forschungsgruppe Wahlen).

Will man Zufriedenheiten messen, *muß* ein Fragebogen genau *so* konstruiert sein wie das 'Studienbarometer'. Gerade die Unmöglichkeit, Fragen 'sachrational' beantworten zu können, zwingt zur Abgabe subjektiver Zufriedenheits-/Unzufriedenheits-Bekundungen. Ihre statistische Zusammenfassung gibt Auskunft über das herrschende Stimmungsbild. Solche Stimmungsbilder können selbstverständlich in einem anderen als dem o.g. Zusammenhang von Interesse und von Nutzen sein: Sie können beispielsweise ein ausgezeichneter Anknüpfungspunkt für Diskussionen über Lehre und Studium an Fachbereichen und

Instituten sein.[2] Der Fehler besteht also nicht in der *Erhebung* von Zufriedenheit, sondern in der *Interpretation* von Zufriedenheit als Qualitätsurteil!

V.
Ich komme zur *zweiten Fehlerquelle* - '**Fehler bei der Informationsaufbereitung**' durch zu einfache Aufbereitung.

Gewählt wird als *Beispiel* für 'einfache Statistik' die häufig anzutreffende Berechnung des arithmetischen Mittels der erhobenen Studentenurteile.

Um an das erste Beispiel - 'Studienbarometer' der FU Berlin - anzuknüpfen: Die befragten Studierenden sind sich natürlich nicht einig in ihren Einschätzungen. Derselbe Sachverhalt wird von einem Teil als ganz schlecht, von einem anderen als mittelmäßig, von noch einem anderen Teil als gut bis sehr gut wahrgenommen. *Das* ließe sich weder in einfacher Weise vermitteln; noch ließe sich so der gewünschte Vergleich zwischen Fachbereichen oder zwischen Erhebungszeitpunkten vornehmen. Also wird auf *einfache Statistik* zurückgegriffen: ein Durchschnittswert muß her (im allgemeinen das aus dem Alltag vertraute arithmetische Mittel[3]): Zu dem *Erhebungs*artefakt (z.B. 'didaktische Kompetenz *der* Lehrenden am Fachbereich') tritt nun zusätzlich ein *statistisches* Artefakt[4] - 'Urteil *der* Studierenden am Fachbereich' - hinzu. Dieses ermöglicht, Urteilsprofile pro Fach grafisch darzustellen und Rangordnungen der Fachbereiche anhand der jeweiligen Einschätzungsaspekte zu bilden.

Nun kann man ganz generell am Informationswert so ermittelter Durchschnitte von ordinalskalierter Daten zweifeln: Was kann es z.B. für das Institut für Soziologie bedeuten , wenn es hinsichtlich der „Hilfsbereitschaft der Verwaltung" den Mittelwert 3,68 erzielt und damit

---

2 Das Projekt pro Lehre der Freien Universität wird zwar nicht müde, bei der Vorstellung der Studienbarometer-Ergebnisse insbesondere auf die Funktion der Anregung von Diskussionen hinzuweisen; aber mit geringem Erfolg. Werden die Befunde rezipiert (etwa in zentralen Gremien der Universität oder in Lehrberichten), dann eben doch im Sinne von „Qualitätsbeurteilungen". Dazu trägt nicht nur der Text des PPL-Begleitschreibens bei, sondern auch die Art der Präsentation der Ergebnisse (darauf wird weiter unten noch einzugehen sein).

3 Dabei wird bei *allen* mir bekannten Darstellungen von Ergebnissen solcher 'Lehrevaluationen' die *Eignung* des arithmetischen Mittels als Instrument der Informationsreduktion nicht einmal ansatzweise diskutiert, sondern - wie im obigen Zitat aus dem Text des Hochschuldidaktischen Zentrums - als im gegebenen Zusammenhang selbstverständlich erfüllt unterstellt.

4 Ein statistisches Artefakt ist das arithmetische Mittel dann, wenn wesentliche Voraussetzungen für die Anwendung des Modells nicht erfüllt sind. Hier trifft das immerhin für zwei Voraussetzungen zu. Zum einen weisen die Daten lediglich Ordinalskalenniveau (statt des meßtheoretisch geforderten metrischen Niveaus) auf: Das ‚+' (linker Skalen-Endpunkt) wird in die Ziffer ‚1' codiert, das ‚–' (reckter Skalen-Endpunkt) in die Ziffer ‚5'; für Zwischenwerte stehen die Ziffern 2, 3 und 4. Eine zweite Voraussetzung liegt in einer Verteilungsannahme des Modells arithmetisches Mittel begründet: Die Meßwerte-Verteilung muß, damit der Mittelwert empirisch sinnvoll interpretiert werden kann, eine *zentrale Tendenz* aufweisen. Diese Voraussetzung ist bei Zufriedenheitsbefragungen häufig empirisch nicht erfüllt (für Lehrveranstaltungsbefragungen s. dazu ausführlicher Kromrey 1995).

auf dem drittletzten Platz unter 32 Fächern rangiert? Was heißt es (und warum sollte es stolz darauf sein), wenn es hinsichtlich der „Praxisnähe des Studiums“ zwar nahezu den gleichen Mittelwert erhält (3,65), aber damit in der obersten Spitzengruppe (nämlich auf Platz 6) liegt? Und ist die „didaktische Kompetenz“ (Mittelwert 2,96 = Rangplatz 13) genauso gut wie die „Arbeitsbedingungen in der Bibliothek“ (2,99 = Platz 20) oder das „Studienklima“ (2,92 = Platz 18)? Und falls man als Institut nicht damit zufrieden sein sollte (es sind ja alles nur mittelmäßige Werte) - was kann man daraus lernen, um die Situation zu verbessern? Die Antwort ist einfach: Nichts kann man daraus lernen; man kann allenfalls Phantasien darüber entwickeln, wer wohl aus welchen Gründen mehr oder weniger zufrieden ist und warum das in anderen Fächern ähnlich oder unterschiedlich ist.

Dieser geringe Informationswert wäre dann nicht besonders problematisch, wenn man sich des tatsächlichen Aussagegehalts des Dargestellten bewußt bliebe, wenn nicht falsche Interpretationen vorgenommen würden (wie sie bewußt in die Darstellung der letzten drei Mittelwerte „hineingeschmuggelt“ wurden): *Nicht Unterschiede in der Qualität der Fächer oder Fachbereiche* werden hier ‘rang-geordnet’, *sondern* in erster Linie *Unterschiede in der Struktur und der Wahrnehmung der befragten Studierenden.* Ob den Unterschieden in den ermittelten Werten auch Unterschiede zwischen den Fächern hinsichtlich der erfragten Sachverhalte entsprechen, läßt sich in keiner Weise (und zwar nicht einmal annähernd) erschließen.

Schon bei ein wenig Nachdenken ist dies leicht nachvollziehbar. Die *zu beurteilenden Sachverhalte* sind im jeweiligen Fachbereich selbstverständlich für alle Befragten *‚objektiv’ identisch*: dasselbe Angebot für alle, dieselben guten oder schlechten Arbeitsbedingungen in der Bibliothek, dieselben vollen oder leeren Lehrveranstaltungen usw. *Subjektiv aber* sind für die einzelnen Befragten die zu beurteilenden Sachverhalte genauso selbstverständlich *nicht identisch*, schon weil die Situation der einzelnen Befragten nicht identisch ist. Die Studierenden unterscheiden sich nach Stellung im Studienverlauf, Haupt- oder Nebenfach, Erst- oder Zweitstudium, Bedeutung des Studiums gegenüber anderen Lebensbereichen, Interesse am Fach, am Inhalt der Veranstaltung, Vorkenntnissen und Erfahrungen, Lernstil, Zeitbudget, Art der Finanzierung des Studiums, Sympathie oder Antipathie gegenüber der Lehrperson, Tagesstimmung und Tagesform, etc. Jede/jeder einzelne steht daher in einer anderen Beziehung zu dem Bewertungsgegenstand, also wird auch ihre/seine Zufriedenheit bzw. Unzufriedenheit unterschiedlich ausfallen müssen:

Je größer die Variation der Situation der Befragten, desto größer ist auch die Variation der Zufriedenheitsurteile über den faktisch identischen Sachverhalt. Und umgekehrt: *Je größer die Antwortvariation, desto heterogener die Zielgruppe* (und desto schwieriger übrigens die Aufgabe der Lehre). Ein Mittelwert mittelt hier nichts aus, sondern schafft das Artefakt einer nicht existierenden Durchschnittszielgruppe. Man ändere die Zusammensetzung der Zielgruppe, und der identische Sachverhalt wird im Durchschnitt anders beurteilt. Wie nun sollte ein einfacher Vergleich *zwischen* verschiedenen Fachbereichen möglich sein, wenn in ihnen nicht nur der *zu beurteilende* Sachverhalt variiert, sondern wenn auch die Situation und die Struktur der beurteilenden Personen variieren?

Ist also - zum Beispiel - die „didaktische Kompetenz“ in der Zahnmedizin (3,00 = Platz 16) wirklich „genauso gut“ wie in der Biologie (3,02 = Platz 21)? Und warum haben sich die Zahnmediziner so „verschlechtert“ (von 2,80 = Platz 7 ein Jahr zuvor)? Weshalb brillieren gerade die Pädagogen sowie das Zentralinstitut für Didaktik mit anscheinend besonders schlechter Didaktik (3,13 = Platz 20 bzw. 3,21 = Platz 27)? Die Antwort auf letztere Frage ist leicht zu vermuten: Die dortigen Studierenden haben ein höheres Anspruchsniveau. Die Antwort auf die vorhergehende Frage fällt ebenfalls leicht: Eine Differenz von 0,2 ist bei einer Stichprobe von n=100 noch gar nicht „statistisch signifikant“ (selbst wenn es sich um ein kontrolliertes Zufallssample gehandelt hätte, was hier allerdings nicht einmal der Fall ist): Es besteht also vermutlich gar kein Unterschied in der Beurteilung, sondern lediglich in der Zusammensetzung der Stichprobe.

Die geschilderte Problematik gilt im übrigen nicht nur für den hier gewählten Fall einer übervereinfachten Datenerhebung. Das wird erkennbar am Beispiel einer differenzierteren Erhebung, in der Pauschalisierungsfehler - wie im obigen Fall - so nicht gelten, nämlich bei der Beurteilung der Lehre jeweils *einer* Lehrperson in *einer* Lehrveranstaltung, an der die Befragten ein halbes Semester teilgenommen haben, mit möglichst trennscharfen Items.[5] Wie gut sind Mittelwerte unter diesen für die Gültigkeit der Einzelurteile vergleichsweise idealen Bedingungen? Die folgende Abbildung illustriert die Situation anschaulich:

[5] In diesem Fall muß also nicht ein subjektiver Mittelwert über mehrere Lehrpersonen anhand pauschaler / zusammenfassender Items gebildet werden; außerdem brauchen die Urteile nicht als ‘spontane’ Schätzungen abgegeben zu werden, sondern fußen auf der wiederholten Erfahrung der Lehre dieser Lehrperson: man hat mehr als ein halbes Semester an der zu beurteilenden Veranstaltung teilgenommen.

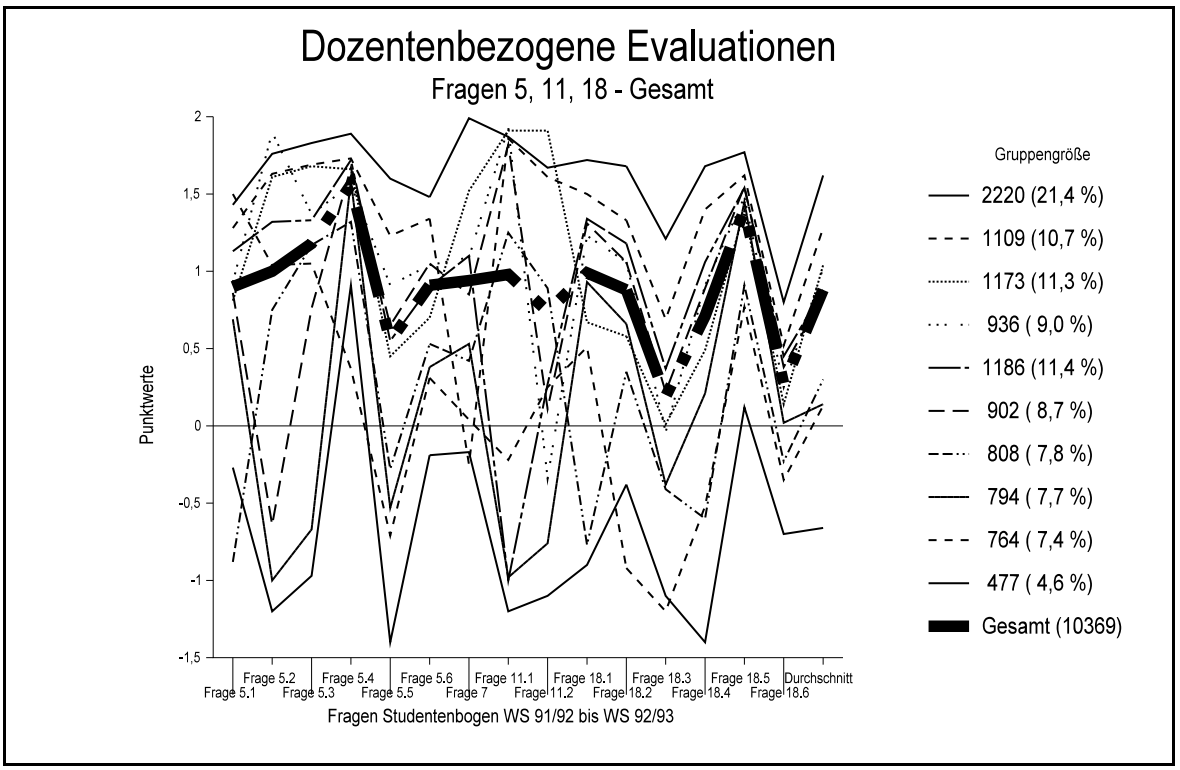

Abbildung 1 Gesamtübersicht (alle Urteilsprofile)

Dargestellt werden in dieser verwirrenden Grafik die Durchschnittswerte der Teilnehmer-Urteile zu einer Reihe von Items (bewertet auf einer 5-Punkte-Ratingskala von -2 bis +2), die sich auf das Lehrverhalten von Dozentinnen und Dozenten beziehen: Vortragsweise, Themenzentrierung, Verständlichkeit, Umfang und Schwierigkeitsgrad des Stoffs, Strukturierung, Klarheit der Lernanforderungen, Dozenten-Interesse, Orientierung an den Studierenden.

Betrachtet man - wie dies oft geschieht - lediglich die Mittelwerte aus den Angaben aller Befragten (breit-gestrichelt gezeichneter Kurvenverlauf), dann könnten die Hochschulen mit dem studentischen Urteil durchaus zufrieden sein: "Im Durchschnitt" wird die Lehre von den Veranstaltungsteilnehmern als "gut" (+1) wahrgenommen - mit zwei Abweichungen ins noch weiter Positive (Themenzentrierung und Grad des Eingehens auf Zwischenfragen: "gerade richtig") sowie ebenfalls zwei Abweichungen hin zum "befriedigend" (± 0) (Klarheit der Lernanforderungen, Berücksichtigung spezieller Wünsche von Studierenden). Entsprechend lauteten erste Reaktionen (auch in der Presse): Die Lehre ist "besser als ihr Ruf".

Bei differenzierterem Blick auf die Daten stellt sich allerdings heraus: Dieses Durchschnittsurteil ist ein statistisches Artefakt, das entsteht, wenn isoliert für die einzelnen Items Mittelwerte berechnet werden. Sucht man mit Hilfe geeigneter statistischer Verfahrern (z.B. Clusteranalyse) danach, wie groß die Gruppe der Studierenden ist, die in dieser Weise "durchschnittlich" urteilt (also ein Urteilsprofil aufweist, das dadurch charakterisiert ist, daß die Dozentenleistung auf fast allen Items in etwa mit +1 bewertet wird), dann zeigt sich: Ein solches "Cluster" existiert nicht. Statt dessen gibt es vielfältige Gruppen von - jeweils

ähnlich urteilenden - Vorlesungsteilnehmern: ganz oben in der Grafik (Abb. 1) die "Fans", denen alles sehr gut erscheint (immerhin mehr als 20 %); ganz unten das entgegengesetzte Extrem: alles erscheint schlecht bis sehr schlecht (wenn auch nur bei knapp 5 %). Dazwischen liegen die Urteilsprofile von Gruppen Studierender, die - in jeweils unterschiedlicher Konstellation - teils positiv, teils negativ werten. Ein statistisch befriedigendes Ergebnis (hinreichend "homogene Cluster", d.h. hinreichend große Gleichartigkeit der Urteile *innerhalb* der jeweiligen Gruppen von Befragten bei zugleich möglichst deutlichen Unterschieden *zwischen* den Gruppen) stellt sich erst ein, wenn mindestens zehn Cluster gebildet werden.

Fazit: Die Uneinigkeit der Studierenden darüber, was als gutes Lehrverhalten empfunden wird, ist außerordentlich groß. Sie besteht darüber hinaus praktisch innerhalb jeder Lehrveranstaltung, d.h. *derselbe* Sachverhalt wirkt positiv auf die einen, negativ auf die anderen. Erfahrene Lehrpersonen werden darüber nicht überrascht sein. Der Befund dürfte mittlerweile auch in der Forschung nicht mehr umstritten sein; er stellt sich auch ein, wenn mit andersartigen Erhebungsverfahren gearbeitet wird.

VI.
Als *dritte Fehlerquelle* hatte ich genannt: '**Fehler bei Schlußfolgerungen und Empfehlungen**'.

Ich will hier nicht auf die naive Vorstellung eingehen, mit einer Verbesserung der Lehre wären alle Probleme des Studierens zu beseitigen; also auf den populären "*Binsenirrtum*", der unterstellt, daß sich eine vorbildliche Ausbildung der Studierenden gleichsam von selbst einstellte, wenn sich die Hochschullehrer bei der Erfüllung ihrer Lehrverpflichtungen nur ausreichend anstrengten.[6] Vielmehr will ich einige *Konsequenzen* aufzeigen, die sich auf den engeren Bereich der Lehre selbst beziehen und die sich *aus methodisch unzulässigen Vereinfachungen bei der Ableitung von Empfehlungen* ergeben.

Zunächst einmal versteht es sich von selbst, daß alle diejenigen Schlußfolgerungen notwendigerweise falsch sind, die sich zu ihrer Begründung auf statistische Artefakte berufen. Wenn schon die empirischen Daten falsch interpretiert werden, können auch die praktischen Ableitungen daraus nicht zutreffend sein.

Darüber hinaus aber können *zwei Typen von Fehlschlüssen* unterlaufen, sofern die Daten selbst zwar keine Artefakte sind, ihre Aufbereitung für den in Frage stehenden Sachverhalt aber zu wenig differenziert vorgenommen wurde. Gerade Sozialwissenschaftler müssen sich tagtäglich mit diesem 'Problem' ihres Gegenstandes auseinandersetzen: Selbst ganz einfach zu formulierende soziale Sachverhalte erweisen sich bei näherem Hinsehen als so komplex, daß sie mit *einfachen* theoretischen Modellen nicht faßbar und dementsprechend auch mit *einfachen* statistischen Kennwerten nicht angemessen darstellbar sind. Das gilt auch für den Gegenstand 'Qualität der Lehre'. Was tun bei diesem Dilemma?

- Ein beliebtes Vorgehen ist es, aus einem insgesamt komplexen *Zusammenhang* einzelne, *isolierte Aspekte herauszugreifen* und Empfehlungen darauf - auf den

---

6 Irrtümer dieses Typs sind auch durch fundierte methodische Ausbildung nicht zu vermeiden; hier ist vielmehr theoretisch fundiertes sozialwissenschaftliches Denken gefragt.

isolierten, einzelnen Aspekt - zu beziehen. In dem hier behandelten Feld “Qualität der Lehre” hört man z.B. von Hochschuldidaktikern häufig Ratschläge von folgendem Typ: Wo Studenten etwas als zu wenig bezeichnen (etwa: zu wenig Medieneinsatz oder zu wenig Beratungsangebote), muß mehr davon geboten werden.

- Entgegengesetzt verfahren Empfehlungen, die die *faktische Differenzierung* des in Frage stehenden Gegenstandes *gänzlich unberücksichtigt lassen* (die also in unserem Fall z.B. pauschal ‘die Studierenden’, ‘die Vorlesungsteilnehmer’ als homogene Gruppe unterstellen).

Wie problematisch beide Vorgehensweisen sind, kann an zwei Ausschnitten aus dem oben (Abb. 1) vorgestellten 'Schnittmusterbogen' - wie ich glaube: eindrucksvoll - illustriert werden. Der erste Ausschnitt (Abb. 2) greift diejenigen Antwortfigurationen heraus, die sich insgesamt in eine Positiv-Negativ-Rangordnung einfügen lassen. Der zweite Ausschnitt (Abb. 3) stellt diejenigen Antwortmuster gegenüber, die 'im Durchschnitt' zu einem gleichen (Gesamt-)Urteil kommen. Die in den beiden Grafiken präsentierten Urteilsprofile sind also neben den inhaltlichen Urteilsunterschieden zugleich Ausdruck von zwei verschiedenartigen Antwortstilen.

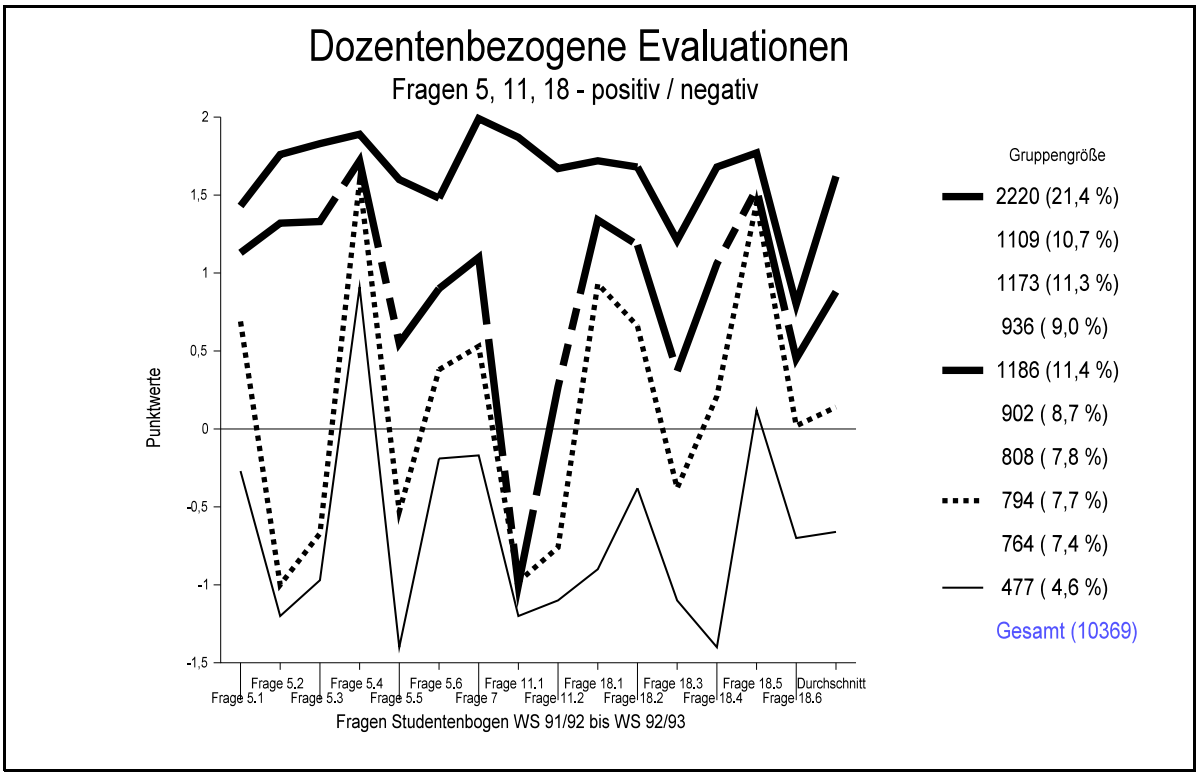

Abbildung 2 Beispiel: rangeordnete Urteilsprofile

Die vier Befragtengruppen, deren Urteilsprofile sich nicht überschneiden (Abb. 2), schätzen die Lehre in der entsprechenden Vorlesung konsistent über alle Detailaspekte hinweg als entweder alles in allem eher positiv oder alles in allem eher negativ ein. Für diesen Teil der

Studierenden (knapp die Hälfte der Vorlesungsteilnehmer) würde die Veränderung eines Einzelaspekts an der Gesamttendenz ihres Urteils nichts (oder zumindest unmerklich wenig) ändern. Die Detailurteile sind Ausdruck der Gesamteinschätzung der Lehrveranstaltung, sie orientieren sich an einer übergreifenden latenten Dimension 'Qualität'. Jede noch so gut gemeinte didaktische Empfehlung, punktuell etwas zu verbessern, bliebe hier wirkungslos.

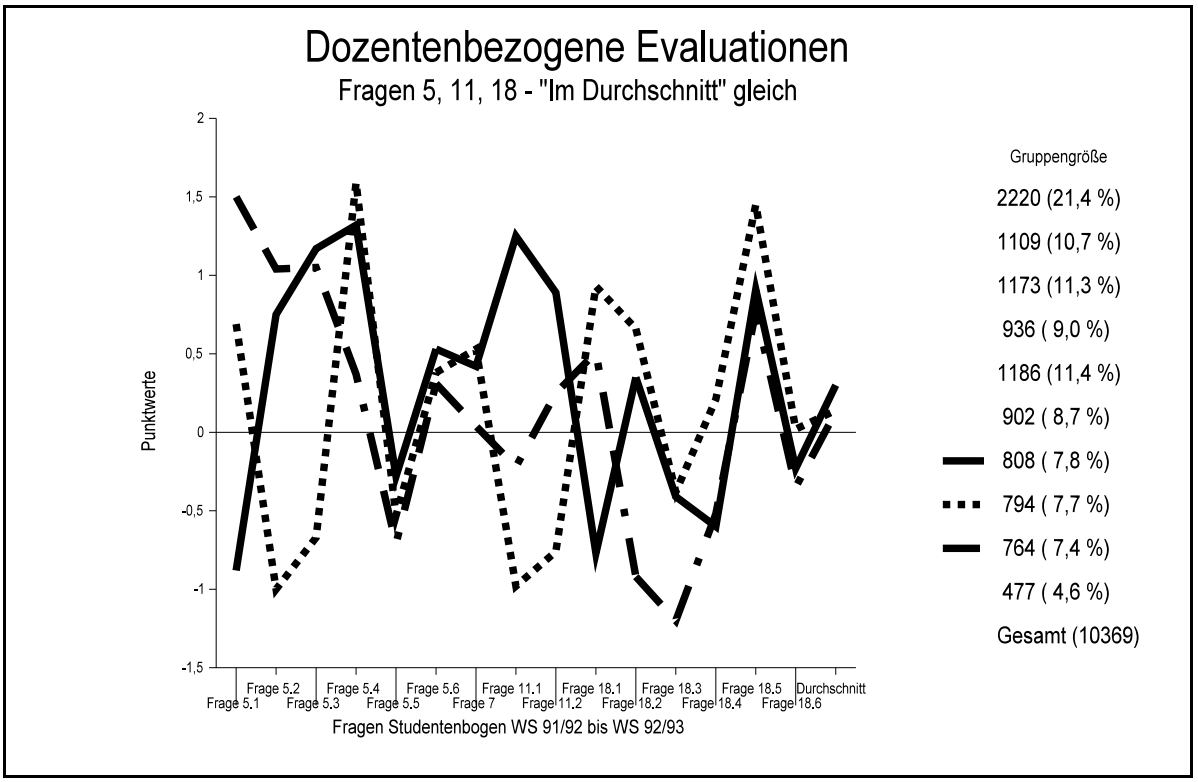

Abbildung 3 Beispiel: Urteilsprofile mit gleichem Durchschnittswert über alle Items

Die auf der zweiten Grafik (Abb. 3) präsentierten Urteilsprofile sind Ausdruck eines entgegengesetzten Antwortstils. Es handelt sich um Studierende, die alles in allem die Lehrveranstaltung gleich-gut oder gleich-schlecht einstufen (nämlich als 'im Durchschnitt mittelmäßig', auch in ihrer eigenen 'Alles-in-allem-Einschätzung'), die jedoch jeden Einzelaspekt differenziert und unabhängig vom Gesamteindruck bewerten. Hier nun führte jede Empfehlung, die 'die Studenten' als eine homogene Gruppe mit gleichgelagerten Ansprüchen an die Lehre betrachtete, vollständig in die Irre.

Da die skizzierten Gruppen von Studierenden mit voneinander abweichenden Ansprüchen an die Lehre (die sich notwendiger- und richtigerweise auch in unterschiedlichen subjektiven Urteilen niederschlagen) *in jeder Lehrveranstaltung in größerer Zahl* vorkommen, gibt es *kein Patentrezept* für gute Lehre. *Jede vereinfachende Empfehlung ist falsch.* Wer versucht, dem Teil der Studierenden gerecht zu werden, der bestimmte Aspekte negativ beurteilt hat, läuft Gefahr, durch eine Veränderung die subjektive Situation derjenigen zu verschlechtern, die bisher positiv eingestellt waren. *Nur zielgruppenbezogene differenzierte Empfehlungen haben überhaupt ein Chance, zutreffend auszufallen.*

VII.

Eigentlich ist es ja erstaunlich, daß eindimensionale Vorstellungen über 'die gute Lehre' immer wieder und immer noch Konjunktur haben; denn zu gleicher Zeit hat selbstverständlich jeder Kenntnis von der Existenz unterschiedlicher Lernertypen und Lernstile. Und wer wollte ernsthaft der Meinung sein, durch eine didaktisch standardisierte Lehre könnten (und sollten) unterschiedliche Lernertypen und Lernstile vereinheitlicht werden? Oder ist dies Ausdruck eines *Theorie-Praxis-Dilemmas der vorherrschenden Variablen-Soziologie*, die in ihren Modellen bestrebt ist, die Effekte jeweils einzelner Wirkungsgrößen herauszuisolieren?

In seiner Antrittsvorlesung als Privatdozent hat *Rolf von Lüde* in Dortmund in diesem Zusammenhang von Paradoxien gesprochen, mit denen sich jeder auseinandersetzen muß, der versucht, seine Lehre studentenorientiert zu optimieren.[7]

So besteht etwa ein '*Lernerfolgs-Paradox*' darin, daß Dozenten in ein unauflösliches Dilemma geraten, wenn sie versuchen, z.B. eine große Vorlesung nach den Kriterien einer studentischen Evaluation zu verbessern, weil einem Hochschullehrer in einer Veranstaltung ganz unterschiedliche Lehr- und Persönlichkeitsprofile zugeschrieben werden. Er wird z.B. in folgende, sich gegenseitig ausschließende Widersprüche verstrickt:[8]

- zu komplex vs. zu trivial (er nimmt uns nicht ernst),
- zu schnell vs. zu langsam,
- zu wenig optische Präsentation vs. Multimediashow usw.
- Und in bezug auf seine Persönlichkeit: zu angepaßt vs. zu exzentrisch.

Ich sprach es eben schon an: Versucht die Lehrperson, der einen Hälfte der Studierenden besser gerecht zu werden, stößt sie notwendigerweise die andere Hälfte vor den Kopf. Das '*Erwartungsdisparitäts-Paradox*' verweist darauf, daß die Bewertung universitäter Lehrqualität stark vom gewählten Qualitätsbegriff bestimmt wird. Legt man einen sowohl „kunden-“ als auch „herstellerorientierten“ Qualitätsbegriff zugrunde und unterscheiden sich die Qualitätserwartungen der Anbieter (also des Lehrpersonals) von denen der nachfragenden Studenten, so kann schon deshalb das Beurteilungsergebnis keine Zustimmung von beiden Seiten finden. Mit anderen Worten: „Rollenbedingte Einschätzungsdisparitäten guter Lehre sind vorprogrammiert und beschreiben den Normalzustand an unseren Hochschulen. Sie verweisen darauf, daß studentische Veranstaltungsbewertungen ausschließlich von einem kundenorientierten Qualitätsmaßstab ausgehen, ..., weshalb theorie- und methodenorientierte Großveranstaltungen häufig in der studentischen Bewertung schlechter abschneiden.“ (a.a.O.)

Schließlich gehen in die Beurteilungen der Qualität von Lehrveranstaltungen Parameter ein, die der Dozent allenfalls indirekt beeinflussen kann, was *von Lüde* mit dem '*Involvement-Paradox*' beschreibt. Versteht man unter Involvement von Studierenden das Eingebundensein in Lehre und Studium, gleichsam das innere Engagement sowie das

---

7 Für eine Kurzfassung s. unizet Nr. 257, 4/1995, S. 4

8 Vgl. hier die Grafiken 'Schnittmusterbogen' (Abb. 1) und 'im Durchschnitt gleich' (Abb. 3).

Arbeitsverhalten, dann ist damit die Frage verbunden, wieviel Energie und Zeit der einzelne bereit ist, in die Mitarbeit in einer Veranstaltung zu investieren. Involvement spielt als externer Faktor für die Qualitätsprüfung einer Vorlesung eine wichtige Rolle - ohne inneres Engagement fällt eine Beurteilung eher negativ aus -, da sie in die Bewertung der Dozentenleistung mit eingeht, ohne von diesem direkt beeinflußbar zu sein. Aufgrund meiner eigenen Erfahrungen läßt sich die hier angesprochene Situation sogar noch weiter zugespitzt formulieren: Unter didaktischen Gesichtspunkten ‚besonders gute' (d.h. ‚nach allen Regeln der didaktischen Kunst' verfahrende) Lehre vergrößert die aus der Medienforschung bekannte ‚Wissenskluft'; sie nützt in erster Linie den ‚guten' (im Idealfall intrinsisch motivierten) Studierenden und ist eher schädlich bei ‚schlechten' (d.h. ohne eigenes Engagement) Studierenden.

IX.
Nach diesem Exkurs abschließend zurück zum Thema ‚Probleme anwendungsorientierter Sozialforschung'. Zu Beginn war ich von der Gefahr ausgegangen, mindestens einen von drei typischen Fehlern zu begehen:

- *Fehler bei der Informationssammlung* durch Orientierung an der Forderung nach Zeit- und Aufwandminimierung, nach Einfachheit und „Laienverträglichkeit"der Methoden,
- *Fehler bei der Informationsaufbereitung* durch Übervereinfachung der Resultate,
- *Fehler bei den Schlußfolgerungen und Empfehlungen* durch Orientierung an *einfachen* Handlungsempfehlungen und an deren leichter Durchsetzbarkeit.

*Soziale Sachverhalte sind immer komplex***.** Im Rahmen von *Grundlagenforschung* oder im Kontext von Theorietest und Theorieentwicklung ist die *Vereinfachung der Untersuchungssituation* (aber auch hier: nicht der Datenerhebung!) sinnvoll und angemessen - bis hin zur Konstruktion völlig realitätsferner Laborsituationen, um den Einfluß eines einzelnen Wirkungsfaktors zu isolieren. Im Rahmen *anwendungsorientierter Forschung* verbietet sich ein solches Vorgehen. *Reale soziale Sachverhalte* sind kein der Vereinfachung zugängliches Experimentierfeld, sondern sind und bleiben komplexe Realität. Eine *angemessene Beschreibung und Diagnose* kann in diesem Rahmen *immer auch nur komplex* ausfallen - oder sie ist zwangsläufig falsch und führt ebenso zwangsläufig zu falschen Schlußfolgerungen.

Grundlagenforschung verfährt nach der Maxime der Wertneutralität. Sie kann und soll sich bei allen zu treffenden Entscheidungen an wissenschaftsimmanenten Normen - und nur an diesen - orientieren (insbesondere Erkenntnisfortschritt als Selbstzweck, Gültigkeit und Zuverlässigkeit der Resultate); außerwissenschaftliche Interessen werden in den Entdeckungs- und Verwertungskontext ausgegliedert. Diese Entlastung von gesellschaftlicher Verwertungs- und Nützlichkeitsargumentation ist gewollt und wesentlicher Bestandteil wissenschaftlicher Legitimation. *Anwendungsorientierte Forschung* muß nicht nur die Definition und Präzisierung ihrer Fragestellungen an außerwissenschaftlichen Erkenntnisinteressen und Verwertungskontexten ausrichten: dies ist schließlich ihr zentrales Definitions- und Legitimationskriterium. Auch bei der methodologischen - also empirisch-wissenschaftlichen - Begründung jeder Detailentscheidung im Forschungsablauf darf die

unmittelbare Verwertbarkeit der erzielbaren Resultate nicht aus dem Blick geraten. „Wissenschaft im Elfenbeinturm“ wird für diese Forschungsrichtung abgelehnt. Ganz im Gegenteil werden wissenschaftsexterne Zweckmäßigkeitsüberlegungen aus dem Verwertungskontext „internalisiert“ und geraten so auf die gleiche Stufe wie die methodologischen Normen der Gültigkeit und Zuverlässigkeit.

Wenn die komplexe Diagnose eines komplexen Zusammenhangs nicht auf Anhieb nachvollziehbar sein sollte, dann liegt die Lösung *nicht* in der *Vereinfachung* (auch unter Inkaufnahme der Gefahr der Verfälschung durch unangemessene Vereinfachung), *sondern* in einer *besseren Erläuterung* durch die Sozialwissenschaftler. Auch dies wiederum ist kein einfaches Unterfangen; es erfordert nicht nur fundierte soziologisch-theoretische Kenntnisse, sondern ebenso fundiertes methodisches Wissen. So ungern manche das hören: Wer die schwierige Disziplin Sozialwissenschaft als Studienfach gewählt hat, muß sich auch den Anforderungen stellen, die zur korrekten Analyse ihres Gegenstandsbereichs unabdingbar sind.

Literatur:

Bundesbildungsminister Ortleb (1992): Thesen zur Belebung der Leistungskraft der Hochschulen. In: Informationen Bildung Wissenschaft, Nr. 7-8/92, S. 90-92.
Kromrey, Helmut (1995): Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozial-forschung, Münster (2. Aufl. 1995): Waxmann, 105-128
MWF - Ministerium für Wissenschaft und Forschung NW (Hg.) (1991): Aktionsprogramm Qualität der Lehre, Düsseldorf
Rindermann, Heiner; Amelang, Manfred (1994): Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE). Handanweisung. Heidelberg: Asanger
Schmidt, Jörn (1980): Evaluation. I. Evaluation als Diagnose, HDZ-Dozentenkurs, Essen (Hochschuldidaktisches Zentrum der Universität Essen GH)