

# Hochschulranking: Beliebigkeit oder konsistente Beurteilungen? Rankings, Expertengruppen und Indikatoren im Vergleich

**Stefan Hornbostel**

Vor mehr als zehn Jahren wurde in Deutschland das erste Hochschulranking veröffentlicht. Damals eher belächelt oder beschimpft, wird es heute, sehr viel ernster genommen (vgl. Hornbostel 1999a). Eine wachsende Zahl von Studienanfängern zieht inzwischen Rankingergebnisse für ihre Studienentscheidung zu Rate, wenngleich die wichtigste Determinante für die Studienortwahl immer noch die Nähe zum heimatlichen Herd ist (vgl. Daniel in diesem Band, Lewin, K. u.a. 1997). Die Rankings selbst haben sich in den vergangenen 10 Jahren erheblich verändert. Dazu muss man nur einmal die jüngste CD des CHE Studienführers mit dem ersten SPIEGEL Ranking vergleichen. Der kritischen Auseinandersetzung mit Rankings kann man eine solche Entwicklung allerdings nicht attestieren. Ein Großteil der Kritik wiederholt stereotyp die Argumente und Verdachtsmoment, die schon vor zehn Jahren vorgebracht wurden.

Eine typische Zusammenfassung der Debatte um Hochschulrankings liest sich folgendermaßen: „Alle sind methodisch bedenklich oder sogar falsch. Die eigentlich propagierten Ziele werden durch die Fragestellungen nur teilweise erreicht. Die Rankingmodelle in Deutschland sind noch immer auf subjektiven Reputationsvergleichen aufgebaut. Im Gegensatz dazu werden die Ranglisten in den USA, die weitaus mehr Tradition als in Deutschland haben, auf Basis von objektiven Kennzahlen erstellt. Die Datenerhebungen selbst beinhalten viele Fehlerquellen wie zu wenig Befragte, falsche Auswahl der Befragten oder fehlerhafte Antworten. Wegen der Unterschiedlichkeit der verwendeten Quellen / Befragten, Fragestellungen, Evaluationskriterien und Bewertungsmaßstäbe sind die Rankings eigentlich nicht miteinander vergleichbar“ (Ott 1999: 320). Diese Liste lässt sich fortsetzen: Gemessen werde mit Hochschulrankings „well being“ sagen die einen (Scheuch 1990), Studierende in den Metropolen seien kritischer als in kleineren Orten, sagen die anderen (Meinefeld 2000) und weisen darauf hin, dass es für Bewertung keine allgemeinverbindlichen Maßstäbe gäbe.

Es herrscht heute zwar weitgehend Einigkeit darüber, dass Evaluation und Ranking notwendig, zumindest aber unumgänglich sind, sobald es aber um tatsächlich durchgeführte Rankingstudien geht, fallen die Einschätzungen über den Ertrag und die Folgen weit auseinander. Während die einen loben, dass Stärken und Schwächen sichtbar, Orientierungsgrundlagen für die Nachfrager hochschulischer Leistungen geschaffen würden, Blockaden und Verkrustungen im Selbststeuerungssystem aufgebrochen oder gar effizienz- und leistungssteigernde Allokationsmechanismen möglich seien, wird von den anderen vorgetragen, dass günstigstenfalls eine Art Infotainment mit einer Fülle widersprüchlicher Informationen entstanden sei oder eine Vergeudung von Zeit und Ressourcen, schlimmstenfalls jedoch ein methodisch fragwürdiges Instrumentarium, das möglicherweise sogar Fehlinformationen und Irreleitungen begünstige oder ein Feigenblatt für weitere Etatkürzungen liefere. Kurz: „Ranking ist oft ein Synonym für Unsinn“ (Süllwoll 1997). Es ist bedauerlich, dass die Auseinandersetzung mit Rankingstudien nach wie vor überwiegend aus Spekulationen und Vermutungen gespeist wird, denn inzwischen liegt eine Fülle von Daten vor, die eigentlich dazu einlädt, sich einmal in konstruktiv-kritischer Weise mit Hochschulrankings auseinander zusetzen.

Der folgende Beitrag versucht, die wichtigsten Einwände gegen Rankingstudien aufzunehmen und mit empirischen Befunden vor allen Dingen aus den CHE Studienführern zu konfrontieren.

### **1. Einwand: Studierendurteile sind eine unzuverlässige Informationsquelle**

Dieses Argument taucht seit Anbeginn der Rankingdebatte in den verschiedensten Versionen auf: es reicht von der Behauptung, dass Studierendurteile allenfalls ein Wohlfühlindikator seien, über die Annahme, Studierendurteile seien vom Aspirationsniveau oder der Eingangsqualifikation der Studierenden abhängig, von Alter, Geschlecht oder der Studienfinanzierung, von den Vergleichsmöglichkeiten mit anderen Hochschulen usw. Im Kern handelt es sich um zwei unterschiedliche Argumente: Einerseits um die Vermutung, Studierende seien zu sachadäquaten Urteilen nicht in der Lage, dabei wird gern darauf verwiesen, dass Professoren häufig gerade jene Hochschulen empfehlen, die von den Studierenden negativ beurteilt werden. Andererseits um ein zweites – damit verwandtes – Argument, das behauptet, dass die *studentischen* Urteile zwar durchaus kompetent seien, aber das Urteil selbst je nach Erwartungen und Voraussetzungen unterschiedlich ausfalle, so dass am Ende nicht die Qualität des Lehrangebotes, sondern Charakteristika der Urteilenden gemessen werden.

Beide Einwände sind empirischer Prüfung zugänglich. Um zu prüfen, ob Studierende überhaupt sachliche Urteile abgeben können, benötigt man eine andere, davon unabhängige Expertise, zum Vergleich. Es liegt nahe, hier die Einschätzungen der Professoren zu Rate zu ziehen, auch wenn beide Gruppen wahrscheinlich eine etwas unterschiedliche Perspektive auf die Lehrsituation haben, schließlich sind die einen Anbieter von Lehre, die anderen Konsumenten. Wären die Kritiken richtig, dass Studierende gerade die hervorragende akademische Ausbildung negativ bewerten und stattdessen nur Nestwärme honorieren, dann müssten die Ranglisten auf den Kopf gestellt werden, wie Scheuch (1990) das einmal formulierte. Da die Professoren sicherlich nicht unter den Verdacht fallen, Nestwärme zu honorieren, müssten die Urteile beider Gruppen negativ korreliert sein.

Wie die Tabelle 1 zeigt, existieren negative Korrelationen zwischen Studierenden- und Professorenurteilen überhaupt nicht, stattdessen zeigt sich eine nach Beurteilungsgegenstand und Fach unterschiedlich hohe Übereinstimmung zwischen beiden Urteilen. Einig sind sich beide Gruppen weitgehend in der Beurteilung relativ einfacher Sachverhalte (wie etwa der Raumsituation). Auf einzelnen Beurteilungsdimensionen und auch bei der Gesamteinschätzung der Studiensituation fallen die Übereinstimmungen fachspezifisch aus. Dabei ist allerdings zu berücksichtigen, dass beiden Gruppen für das Gesamturteil eine unterschiedliche Frage vorlag. Auch die Art der Übereinstimmung lässt fachspezifische Differenzen erkennen: In der Rechtswissenschaft fallen die Urteile in der gleichen Richtung aus, aber auf unterschiedlichen Niveaus (die Professoren vergeben bessere Noten als die Studierenden), in der Physik hingegen stimmen die vergebenen Noten in hohem Maße überein und bei den Architekten herrscht vergleichsweise geringe Übereinstimmung zwischen Professoren und Studierenden. Insgesamt lässt sich jedoch festhalten, dass größerer Dissens zwischen beiden Gruppen nicht die Regel ist, sondern nur an wenigen Hochschulen auftritt. Trotz der unterschiedlichen Perspektiven von Produzenten und Konsumenten kommen beide Gruppen in der Tendenz zu ähnlichen Beurteilungen.

	Rechtswissen.	Physik	Informatik	Bauing.	Masch. bau	Mathematik	Elektrotech.	Architektur
Index Räume	.80**	.77**	.75**	.70**	.77**	.66**	.55**	.55*
Index Bibliothek	.58**	.46**	.47**	.63**	.44*	.20	.43*	.45
Kontakt Stud.-Prof.	.55**	.60**	.28	.57*	.48*	.36*	.66**	.17
Index PC	.46**	.44**	.52**	.11	.39*	.55**	.50*	.46
Gesamturteil Lehre <sup>1</sup>	.51**	.42**	.40**	.52*	.35	.36*	.27	.24

**Tab. 1: Korrelationen (Pearson) zwischen den Fachbereichsmittelwerten der Beurteilungen von Professoren und Studierende. Quelle: CHE Studienführer 1999 und 2000.**

\* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

\*\* Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

**Fazit: Berücksichtigt man, dass auch ausgewiesene Experten (etwa Gutachter bei der Beurteilung von Manuskripten oder Drittmittelanträgen) erheblichen Urteilsdissens aufweisen, dann deutet das Maß an Übereinstimmung zwischen Professoren- und Studierendenurteilen auf hohe Urteilskompetenz bei den Studierenden hin.**

## **2. Einwand: Professoren empfehlen gerade die Hochschulen, die von Studierenden schlecht beurteilt wurden**

In der Berichterstattung über Rankingstudien wird immer wieder darauf hingewiesen, dass Professoren andere Hochschulen empfehlen als die von den Studierenden positiv beurteilten. Wie ist eine solche Diskrepanz zu erklären?

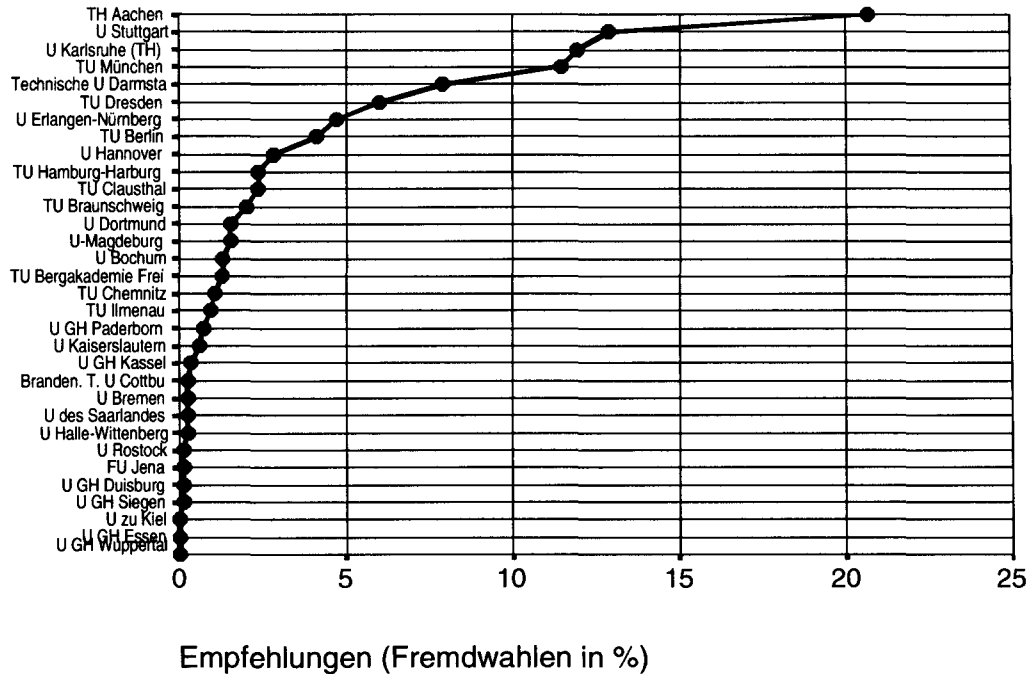
Für den obigen Vergleich wurden die Professoren gebeten, die Lehrsituation vor Ort zu beurteilen. Daneben wurden in der CHE Befragung – wie auch in anderen Rankingstudien – die Professoren um eine Empfehlung, einen „Studententipp“ gebeten (also um eine Fremdeinschätzung; vgl. dazu den Beitrag von Daniel in diesem Band). Während die ersteren Professorenauskünfte sich auf die eigenen Erfahrungen an der derzeitigen Hochschule stützen, basieren die Empfehlungen auf einer Beobachtung der Kommunikation über universitäre Leistungen. Kommuniziert werden im Wissenschaftssystem aber nicht Lehr-, sondern Forschungsleistungen (vgl. ausführlich Hornbostel 2001a).

Die Studienortempfehlungen der Professoren haben in allen Fächern die Form einer sehr schiefen und steilen Verteilung. Renommee und Bekanntheit konzentrieren sich meist auf wenige Hochschulen. Diese Verteilung ist unabhängig davon, ob „Selbstempfehlungen“ der eigenen Universität berücksichtigt oder aus der Wertung ausgeschlossen werden. Abbildung 1 zeigt einen solchen typischen Verlauf des „Professorentipps“.

<sup>1</sup> Die Frage für die Studierenden lautete: „Wenn Sie nun einmal alles zusammen betrachten: Wie beurteilen Sie insgesamt die Studiensituation in Ihrem Fach an Ihrer derzeitigen Hochschule?“

Die Frage für die Professoren lautete: „Wie beurteilen Sie insgesamt die Lehrsituation in Ihrem Fach an Ihrer derzeitigen Hochschule?“

Abb. 1: Hochschultipp der Professoren  
Maschinenbau / Verfahrenstechnik (1999)



Für die Fremdwahrnehmung einer Hochschule bzw. eines Fachbereiches ist die Quantität und die Qualität des Forschungsoutputs wesentlich. Tabelle 2 zeigt diesen Sachverhalt recht deutlich. Sowohl die verschiedenen Größenangaben als auch die diversen Forschungsindikatoren korrelieren positiv mit der Zahl der abgegebenen Empfehlungen für eine Hochschule. Dass dabei in der Rechtswissenschaft die Drittmittelwerbungen keine signifikanten Korrelationen aufweisen, hat mit der extrem niedrigen Drittmittelaktivität in diesem Fach zu tun und mit dem Umstand, dass die Bewertungskriterien für Forschung unschärfer sind und die adressierbaren Auditorien nicht nur wissenschaftsinterner Art sind, sondern auch in die Rechtspraxis ausgreifen (vgl. Hornbostel 1999).

Betrachtet man nun die subjektiven Urteile zunächst der Professoren, zeigt sich, dass an den besonders stark empfohlenen Hochschulen die Forschungssituation vor Ort tendenziell besser bewertet wird, die Beurteilung der Lehrsituation hingegen korreliert weitaus schwächer und nicht signifikant mit den Empfehlungen der Professoren im Fach<sup>2</sup>. Vergleicht man diese Beurteilung der Professoren mit den studentischen Bewertungen, lassen sich keine gravierenden Unterschiede ausmachen. Ebenso wie die Studierenden bewerten die Professoren vor Ort die Lehrsituation weitaus kritischer, als die Kollegen anderer Universitäten mit ihren Studienortempfehlungen zum Ausdruck bringen.

<sup>2</sup> Die negative Korrelation entsteht, weil hier gute Benotungen (niedrige Zahlenwerte) mit einem hohen Prozentsatz der abgegebenen Empfehlungen (hohe Zahlenwerte) korrespondieren.

	Informatik	Mathematik	Physik	Jura
<b>Struktur</b>				
Studierende am Fachbereich	.43 **	.43 **	.35 **	.41 **
Studierende an der Universität gesamt	.02	.30 *	.35 **	.37 *
Prüfungen je Prof. im Fach (1997)	.52 **	.23	.34 *	.47 **
<b>Forschung</b>				
Promotionen je Prof. im Fach (1997)	.46 **	.29 *	.36 **	.64 **
Drittmittel (1996-98) absolut	.82 **	.26 *	.58 **	.21
Drittmittel (1996-98) je Wissenschaftler	.59 **	.36 **	.25	-.04
Artikel 1996-98 im SCI (absolut)		.49 **	.73 **	
Artikel 1996-98 im SCI je Wissenschaftler		.41 **	.47 **	
Zitate je Publikation			.36 **	
<b>Urteile</b>				
Gesamtbeurteilung Forschungssituation (Prof.)	-.41 *	-.36 **	-.32 *	-.06
Gesamtbeurteilung Lehrsituation (Profess.)	-.22	-.20	-.08	.08
Gesamtbeurteilung Studiensituation (Studierend.)	-.01	.01	.10	.11

Tab. 2: Korrelationen (Pearson) zwischen Hochschulempfehlungen der Professoren (Prozent der Nennungen ohne Selbstwahlen) und Strukturindikatoren sowie Beurteilungen von Professoren und Studierenden

**Fazit: Die Differenz zwischen Studienortempfehlung der Professoren und den Bewertungen der Lehr- und Studiensituation vor Ort entspricht einer Differenz von Fremd- und Selbstbild, wobei das Fremdbild wesentlich über die Forschungsleistungen geprägt wird. Studierende urteilen dabei nicht anders als Professoren auch. Die Studienempfehlungen der Professoren indizieren Forschungsperformanz.**

Auch die Frage, ob das studentische Urteil durch Merkmale der Studierenden verzerrt wird, ist immer wieder empirisch überprüft worden (z. B. Daniel / Hornbostel 1994 und 1996). Die Daten des CHE Studienführers bestätigen zunächst einmal die bisherigen Ergebnisse. Wie Tabelle 3 zeigt, geht der ganz überwiegende Teil der aufgeklärten Varianz im studentischen Gesamturteil auf die Zugehörigkeit zu einem Fachbereich zurück und nicht auf individuelle Merkmale der Studierenden. Geschlecht und Studiendauer spielen überhaupt keine Rolle für die Beurteilung der Studiensituation. Das Lebensalter der Studierenden wirkt sich nur geringfügig aus. Das Eingangsniveau der Studierenden (hier über die Abiturnote erfasst) beeinflusst nur in der Rechtswissenschaft und der Elektrotechnik geringfügig das Urteil. Studierende mit schlechten Schulabgangsnoten beurteilen ihre Studiensituation etwas schlechter (dichotomisiert man die

Population in Schulnote  $\leq 2$  und  $> 2$ , dann beträgt die mittlere Differenz in der Gesamtbeurteilung der Studiensituation in der Rechtswissenschaft allerdings nur 0,26 Skaleneinheiten).

	Maschbau	Elektrotechnik	Bauing	Architektur	Informatik	Mathematik	Physik	Rechtswiss.
Fachbereich	.72	.53	.64	.80	.72	.55	.69	.66
Forschungs-/ Praxisorientiert <sup>3</sup>	.07	.11	.14	.02	.13	.21	.07	.07
Vollzeit- / Teilzeitstudierend <sup>4</sup>	.13	.11	.11	.03	.11	.07	.09	.07
Alter	.00	.14	.01	.04	.00	.14	.08	.02
Geschlecht	.01	.00	.00	.05	.00	.00	.00	.03
Abiturnote	.04	.10	.08	.00	.03	.03	.04	.11
Job	.01	.00	.00	.00	.01	.00	.01	.00
Fachsemester	.03	.02	.03	.06	.00	.01	.02	.04
korrigiertes R <sup>2</sup>	.19	.24	.23	.27	.21	.21	.22	.29
Valide Fälle	1931	1270	1270	957	1037	1293	1211	1307

Tab. 3: Regression (CATREG) auf das studentische Gesamturteil (abhängige Variable) - relative Wichtigkeit nach Pratt und korrigiertes R<sup>2</sup>. Quelle: CHE Studienführer 1999 und 2000.

In den Ingenieurwissenschaften, der Informatik und der Physik wird, wenn man vom Einfluss des Hochschulstandorts absieht, das Urteil geringfügig vom zeitlichen Engagement der Studierenden beeinflusst. Die ca. 15% Physikstudierenden, die sich selbst als „Teilzeitstudierende“ einstufen, obwohl dieser Status de jure nicht existiert, beurteilen ihre Studiensituation insgesamt auf der Schulnotenskala von 1 bis 6 fast einen halben Punkt schlechter als ihre Kommilitonen. Auch die etwa 27% Informatikstudierenden, die sich als Teilzeitstudierende definieren, urteilen in dieser Größenordnung schlechter als ihre Kommilitonen. In diesen Fächern bieten sich für Studierende bereits im Studium attraktive Möglichkeiten auf dem Arbeitsmarkt. Möglicherweise verschiebt ein quasi-berufliches Engagement die Erwartung an Inhalte und Effizienz des Studiums. Allerdings in einer Größenordnung, die – würde man diesen Effekt herauspartialisieren – die Positionierung der Fachbereiche fast unverändert lässt.

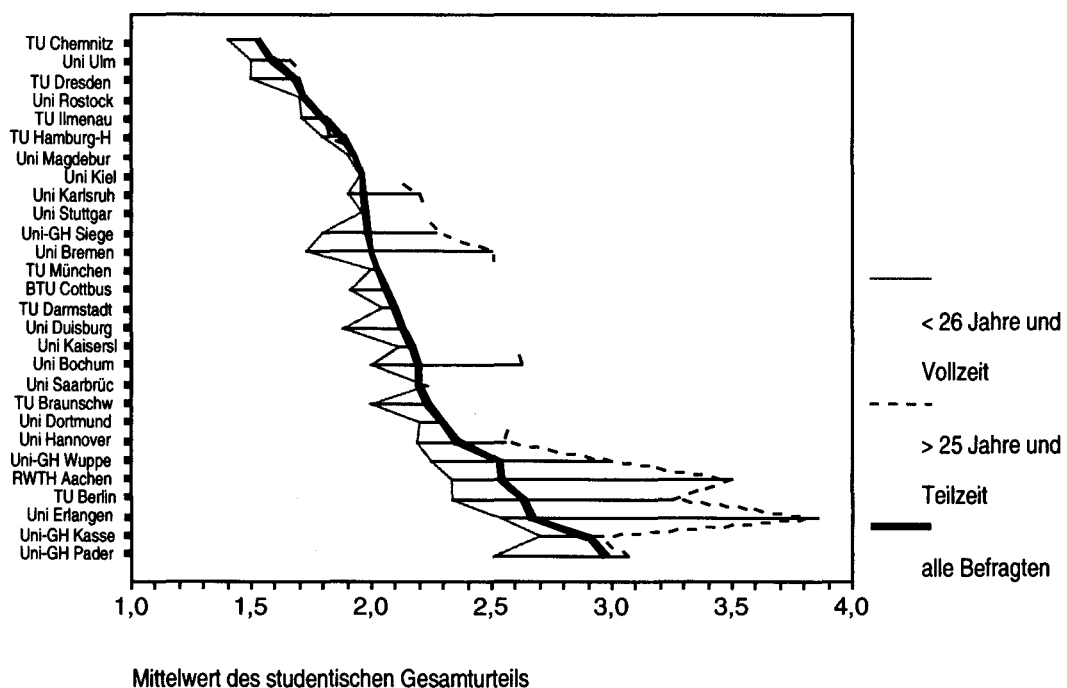
3 Aus den Fragen nach der Wichtigkeit des Forschungsbezuges bzw. des Praxisbezuges des Studiums (6 stufige Skala) wurden drei Typen gebildet: 1. Studierende, für die der Forschungsbezug wichtig ist (Skaleneinheiten 1 und 2) und der Praxisbezug gleich oder weniger wichtig ist; 2. Studierende, für die der Praxisbezug wichtig ist (Skaleneinheiten 1 und 2) und der Forschungsbezug weniger wichtig ist; 3. Studierende, für die sowohl der Forschungsbezug als auch der Praxisbezug weniger wichtig ist (Skaleneinheiten  $> 2$ ).

4 Die Angaben beruhen auf einer Selbsteinstufung als Vollzeit oder Teilzeitstudierend, unabhängig vom Vorhandensein einer entsprechenden rechtlichen Regelung.

Von den 1.565 befragten Studierenden der Elektrotechnik gehören ca. 66% zum „Normaltypus“ (jünger als 27 Jahre und Vollzeitstudierend) und ca. 11% zum in der Regel besonders kritisch urteilendem Typus, der älter als 26 Jahre ist und sich als teilzeitstudierend einstuft. Letztere Studierenden urteilen im Durchschnitt ca. 0,7 Skaleneinheiten schlechter (Gesamturteil) als die Studierenden aus der ersten Gruppe. Der Anteil der älteren Teilzeitstudierenden an allen Studierenden der einzelnen Fachbereiche schwankt zwischen ca. 2% und ca. 32%. Die Existenz solcher Subgruppen wird häufig als Argument gegen die Aussagekraft von Mittelwerten (vgl. Süllwold 1997) verwandt, da sich im Extremfall ähnlich große Gruppen mit diametral entgegengesetzten Urteilen an einem Fachbereich befinden. Praktisch sieht es allerdings meist anders aus, wie die folgende Abbildung 2 zeigt. In der Graphik werden die Urteilsmittelwerte der jungen Vollzeitstudierenden, der älteren Teilzeitstudierenden und aller Studierenden in der Elektrotechnik verglichen (berücksichtigt sind nur die Fälle, in denen in den Subgruppen fünf oder mehr Studierenden vertreten sind).

## Elektro- und Informationstechnik

### Gesamturteil nach Subgruppen



Erkennbar ist zunächst, dass zwischen beiden Subgruppen tatsächlich erhebliche Urteilsunterschiede bestehen. Allerdings zeigt die Graphik auch, dass beide Gruppen zu ähnlichen relativen Beurteilungen (verglichen mit den übrigen Fachbereichen) kommen, d.h. keiner der schlecht beurteilten Fachbereiche würde in die Spitzengruppe aufrücken, wenn man nur das Urteil der jüngeren Vollzeitstudierenden heranzieht. Ähnliches gilt für das Urteil der älteren Teilzeitstudierenden: Die Positionierung der Fachbereiche ändert sich nur geringfügig, wenn man ausschließlich das Urteil der älteren Studierenden berücksichtigt. Wie Abbildung 2 zeigt, ist trotz erheblicher Bewertungsunterschiede der Mittelwert aus allen studentischen Urteilen ein guter Schätzer für die Positionierung der Fachbereiche.

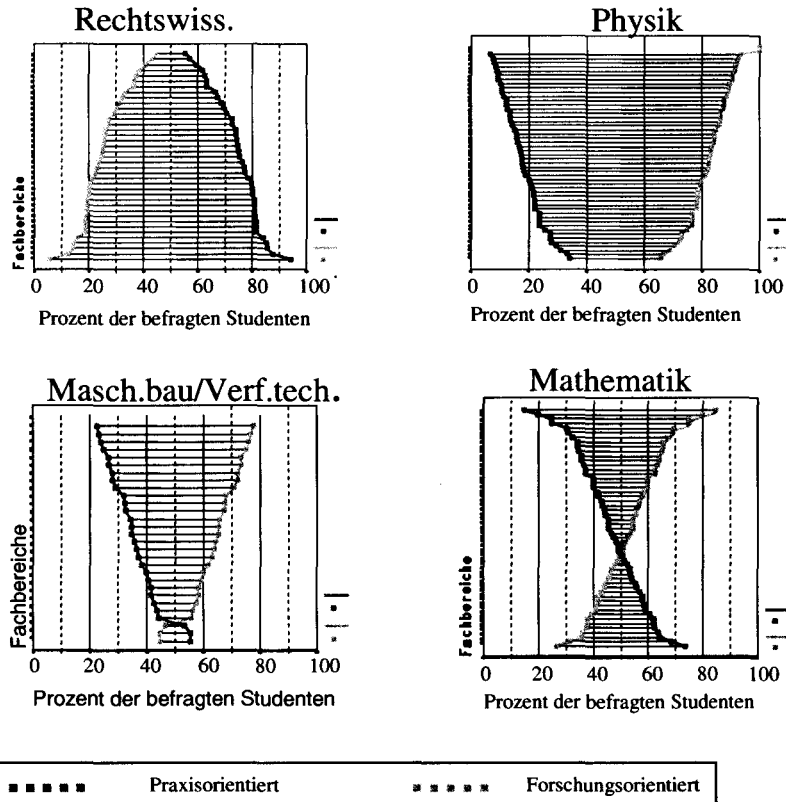
Die Existenz von Subgruppen mit unterschiedlichen Urteilsniveaus an sich ist daher kein Problem für eine grobe Positionierung der Fachbereiche. Dies ändert sich erst, wenn extrem inhomogene Fachkulturen vorliegen, d.h. wenn Subgruppen nicht nur völlig entgegengesetzt werten, sondern ihr Anteil an der jeweiligen Studierendenpopulation auch extrem schwankt.

### Abb. 3: Wichtigkeit der Forschungs- und Praxisorientierung im Studium

Bewertungen der Studierenden verschiedener Disziplinen auf einer 6 stufigen Skala.

Praxisorientierung = % der Befragten, die den Praxisbezug für wichtig halten (Skalenwerte und 1 & 2) und den Forschungsbezug für weniger wichtig halten oder Praxis- und Forschungsbezug für weniger wichtig halten.

Forschungsorientierung = % der Befragten, die den Forschungsbezug für wichtig halten (Skalenwerte und 1 & 2) und den Praxisbezug für weniger wichtig halten oder Praxis- und Forschungsbezug für wichtig halten.



Derartige Phänomene lassen sich im Fach Mathematik beobachten: Zu den auffälligsten Einflussfaktoren in Tabelle 2 gehört die Einschätzung der Studierenden, ob eher der Forschungs- oder eher der Praxisbezug des Studiums wichtig sei. Von den Studierenden der Rechtswissenschaft halten 54% den Praxisbezug für wichtiger, von den Informatikstudierenden immerhin noch 28%. In der Physik hingegen halten 19% der Studierenden den Forschungsbezug für wichtiger, und die große Mehrheit stuft beide Aspekte gleichermaßen als sehr wichtig oder wichtig ein. Abweichend von diesem Muster finden sich in der Mathematik zwei Teilpopulationen von 16% und 24 % der Befragten, die entweder der Praxis oder der Forschung höheren Stellenwert einräumen. Diese Dichotomie reflektiert den enormen Bedeutungszuwachs, den die angewandte Mathematik neben der reinen Grundlagenmathematik erhalten hat. Während in der Rechtswissenschaft und in der Physik jeweils die Forschungs- bzw. Praxisorientierung klar dominiert, finden sich in der Mathematik sowohl Fachbereiche, in denen die Praxisorientierung

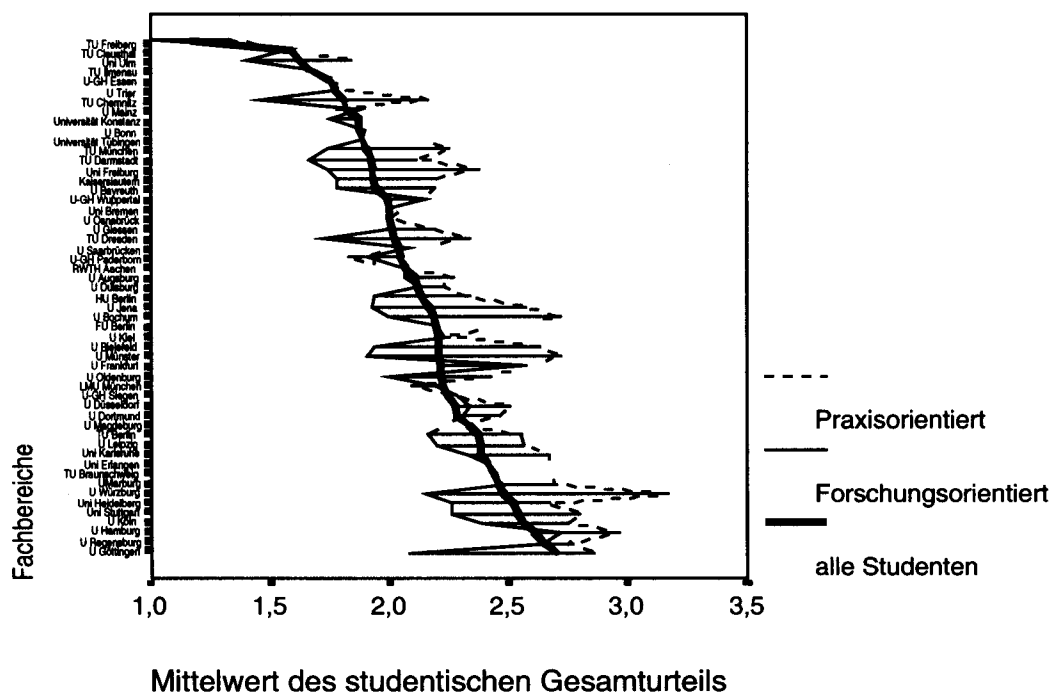
überwiegt, als auch Fachbereiche, in denen die Forschungsorientierung überwiegt<sup>5</sup> (vgl. Abb. 3). Auch in der Mathematik vergeben die praxisorientierten Studierenden in der Regel schlechtere Noten. Vergleicht man nun die Urteilsittelwerte von praxisorientierten Studierenden mit denen der forschungsorientierten an den einzelnen Fachbereichen, dann ergibt sich zwar immer noch eine signifikant positive Rangkorrelation zwischen beiden Gruppen (Kendalls Tau  $b = .47$ ), aber die Zahl der Fachbereiche, die ihre Rangposition erheblich ändern, ist relativ groß: Rund 40% der vergleichbaren 38 Fachbereiche ändert die Rangposition um zehn oder mehr Plätze, wenn man die Positionierungen einmal aufgrund der Bewertungen der forschungsorientierten Studierenden durchführt und einmal auf Grund der Beurteilung der praxisorientierten Studierenden.

Die Zufriedenheit mit den Studienbedingungen hängt offenbar davon ab, ob es gelingt, den unterschiedlichen Erwartungshaltungen gerecht zu werden, die mit einer eher praktischen bzw. einer eher theoretisch-forschungsbezogenen Orientierung verbunden sind.

Abb. 4

## Mathematik

### Gesamturteil nach Subgruppen



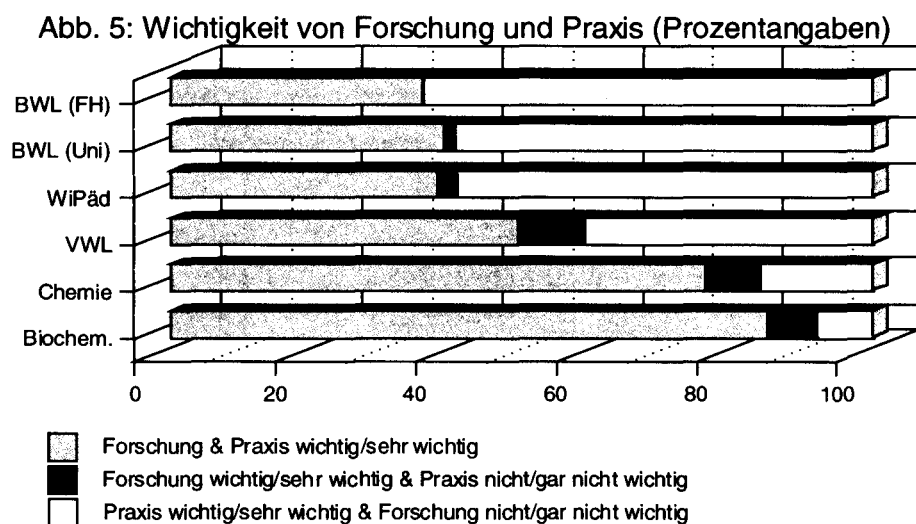
Solche dramatischen Positionsveränderungen ergeben sich, weil die beiden Subgruppen sich nicht einfach durch eine Niveaushiftung in ihrem Urteil auszeichnen, sondern durch eine fachbereichstypische Differenz in den Urteilen. An rund 40% der Fachbereiche liegen die Urteile beider Gruppen nur bis zu 0,3 Skalenpunkte auseinander, aber an rund 24% der Fachberei-

<sup>5</sup> Dabei handelt es sich bei den hier wiedergegebenen CHE Untersuchungen ausschließlich um Studierenden mit dem Abschlussziel „Diplom“, die Berücksichtigung von Lehramtsstudierenden würde das Problem vermutlich noch mehr verschärfen.

che beträgt die Differenz 0,6 und mehr Skalenspunkte (bezogen auf eine sechsstufige Antwortskala). Diese besonderen Verhältnisse in der Mathematik erklären auch, warum unterschiedliche Rankingstudien in diesem Fach zu unterschiedlichen Ergebnissen kommen.

**Fazit: Die bisherigen empirischen Befunde deuten darauf hin, dass bias-Variablen einen allenfalls geringen Einfluss auf das Urteil der Studierenden haben und entsprechend bei einer Korrektur nur geringfügige Positionsveränderungen der Fachbereiche in Ranglisten zur Folge haben. Ausnahmen davon treten offenbar nur unter besonderen, fachspezifischen Bedingungen (vgl. Mathematik) auf. Ein Generalvorbehalt ist gegenüber der Aussagefähigkeit studentischer Beurteilungen ebenso ungerechtfertigt, wie gegenüber der Verwendung von Mittelwerten zur Charakterisierung der Fachbereiche.**

Derartige Phänomene, wie sie in der Mathematik auftreten, werfen natürlich die Frage auf, ob die erfragten Sachverhalte aus der Sicht der Studierenden überhaupt relevant für die Qualität eines Studienangebotes sind. Folgt man den durch Marketingüberlegungen inspirierten Konzepten unterschiedlicher Rankingzielgruppen (vgl. Bayer in diesem Band), dann müsste jede Frage zusätzlich mit einer Relevanzeinschätzung verbunden werden. Items mit hoher Relevanz aus der Sicht der Studierenden erhalten dann ein größeres Gewicht als Items mit geringer Relevanz. Ein solches Vorgehen wurde im ersten Studienführer des CHE / Stiftung Warentest (1998) gewählt. Neben der Bewertung einzelner Sachverhalte wurden die Studierenden gebeten, auf einer vierstufigen Skala (sehr wichtig bis gar nicht wichtig) jedes Item mit einer Relevanzeinstufung zu versehen.



**Anm: Bewertung der Wichtigkeit des Forschungs- bzw. Praxisbezuges der Lehrveranstaltungen auf einer vierstufigen Skala von „sehr wichtig“ bis „gar nicht wichtig“. Quelle: CHE / Stiftung Warentest Studienbefragung 1998**

Greift man zunächst einmal den oben schon dargestellten Bereich von Forschungs- und Praxisbezug der Lehrveranstaltungen heraus, dann zeigt sich wiederum eine erhebliche Differenz zwischen anwendungs- und forschungsorientierten Disziplinen: Während in der Betriebswirtschaft die Studierenden, an Fachhochschulen ebenso wie an den Universitäten, überwiegend den Pra-

xisbezug für wichtiger halten als den Forschungsbezug, werden in der Chemie und Biochemie überwiegend Praxis- und Forschungsbezug für gleichermaßen wichtig gehalten, zugleich finden sich zwei kleinere Teilpopulationen mit expliziter Forschungs- bzw. Praxisorientierung. Wenn diese Orientierungen zu unterschiedlichen Erwartungshaltungen an das Studium und in der Folge zu unterschiedlichen Bewertungen führen, so könnten sich auch die Mittelwerte der studentischen Beurteilungen auf Fachbereichsebene verändern, wenn man die unterschiedlichen Relevanzeinschätzungen der Studierenden berücksichtigt. Bei solchen Überlegungen muss allerdings sehr genau zwischen Effekten auf der Individualebene und solchen auf der Aggregatebene unterschieden werden. Auf der Aggregatebene wirkt sich eine Berücksichtigung der Relevanzeinschätzungen nur aus, wenn a) Subgruppen existieren, die sich in der Beurteilung und der Relevanzeinschätzung deutlich von ihren Kommilitonen unterscheiden, und b) wenn diese Subgruppen nicht in allen Fachbereichen mehr oder weniger gleichmäßig auftauchen, sondern sich an einzelnen Standorten konzentrieren. Nur unter diesen Bedingungen würde sich die Rangfolge der Fachbereiche tatsächlich ändern. Auf der Ebene von Gedankenexperimenten werden derartige Überlegungen gern als Beleg für die fehlende Aussagefähigkeit von Urteilsmitteiwerten angeführt.

Da beide Bedingungen aber in aller Regel jedoch nicht erfüllt sind, ergibt eine Berücksichtigung der studentischen Wichtigkeitseinschätzungen empirisch keine Veränderung der Positionierung der Fachbereiche. Dieser Befund ist unabhängig davon, welche Gewichtungsverfahren benutzt werden.

	Wichtigkeit als Gewichtungvariable	Multiplikative Gewichtung der Urteile	N
Inhaltliche Breite des Lehrangebotes	0,998 (p = 0,000)	0,902 (p = 0,000)	48
Forschungsbezug der Lehrveranstaltungen	0,993 (p = 0,000)	0,953 (p = 0,000)	48
Verfügbarkeit der benötigten Literatur	1,000 (p = 0,000)	0,964 (p = 0,000)	48
Ausstattung der Computerarbeitsplätze	0,998 (p = 0,000)	0,951 (p = 0,000)	48

**Tab. 4: Korrelation (Pearson) zwischen gewichteten und ungewichteten Beurteilungen (Fachbereichsmittelwerte- BWL -Universitäten). Quelle: Studierendenbefragung CHE / Stiftung Warentest 1998**

Tabelle 4 zeigt für verschiedene Fragekomplexe die Korrelationskoeffizienten zwischen den ungewichteten Mittelwerten der Fachbereiche und einer Gewichtung durch Mehrfachberücksichtigung der Fälle (Wichtigkeit als Gewichtungvariable) sowie einer multiplikativen Gewichtung der Beurteilung mit der Wichtigkeitseinschätzung. In beiden Fällen ergeben sich nahezu perfekte Korrelationsbeziehungen. Aufgrund dieser Befunde wurde in den folgenden Studienführern auf die Erhebung von Wichtigkeitsurteilen zu jedem Item verzichtet.

**Fazit: Die Gewichtung von Urteilen mit den Relevanzeinschätzungen des beurteilten Sachverhaltes hat auf der Aggregatebene (Fachbereich) kaum Auswirkungen.**

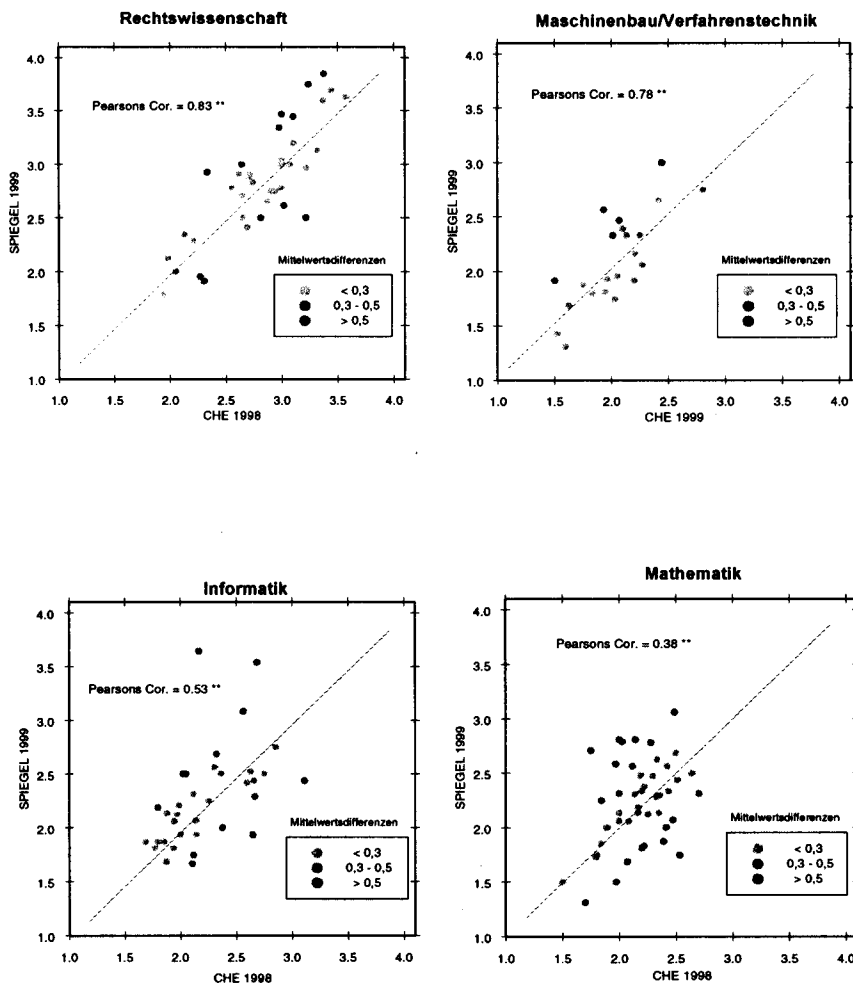
### **3. Einwand: Rankings führen zu instabilen, fast beliebigen Ergebnissen**

Dieser Einwand beruht fast regelmäßig darauf, dass die an anderer Stelle eingeklagte Differenzierung von Rankingergebnissen beim Vergleich unterschiedlicher Studien völlig vernachlässigt wird. Ernsthaft wird wohl niemand erwarten, dass die Befragung unterschiedlicher Auskunftspersonen (von Managern über Professoren zu Studierenden), die Verfolgung unterschiedlicher Fragestellungen (von Imagestudien über indikatorengestützte Bewertungen bis hin zu subjektiven Einschätzungen) und unterschiedliche Verrechnungen von Einzelbefunden zu identischen Ergebnissen führen.

Vergleichen lassen sich nur Studien mit zumindest ähnlichem Design. Ein solcher Vergleich eröffnet zugleich die Möglichkeit die Reliabilität des Instrumentes bzw. den Einfluss verzerrter Stichproben zu prüfen. Die SPIEGEL Erhebung von 1999 und die CHE Studienführer 1999 und 2000 enthalten in der Studierendenbefragung ein ähnliches Item, in dem die Studierenden um eine Gesamtbeurteilung der Studiensituation gebeten werden. Ansonsten unterscheiden sich die Studien erheblich: Die SPIEGEL Befragung erfasst Studierenden aller Abschlussarten, die CHE Studie nur Diplom bzw. Staatsexamen; die SPIEGEL Studie basiert auf einer willkürlichen Auswahl der Studierenden mit Ortsvorgaben für die Interviewer, die CHE Studien beruhen größenabhängig auf Vollerhebungen oder auf Zufallsstichproben, die SPIEGEL Umfrage ist als mündliche Campusbefragung durchgeführt worden, die CHE Studien als schriftliche, postalische Befragung, schließlich wurden die CHE Erhebungen mit einem weitaus größerem Stichprobenumfang durchgeführt und last not least: die Studien wurden zu unterschiedlichen Zeitpunkten durchgeführt, so dass Reaktanzen nicht ausgeschlossen werden können. Unterschiede also, die bei Methodenvergleichen in der Regel bereits deutliche Abweichungen in den Ergebnissen produzieren. Vergleicht man nun die Fachbereichsmittelwerte aus beiden Studien, dann zeigen sich für die Rechtswissenschaft und den Maschinenbau sehr hohe Übereinstimmungen (Pearson Cor. .78 bzw. .83; vgl. Abb. 6). Dieser Befund gilt auch für die Informatik mit der Einschränkung, dass hier zwei extreme Ausreißer den Korrelationskoeffizienten (.53) „drücken“. In der Mathematik hingegen ist zwar immer noch eine deutliche Korrelation zu erkennen, aber bei einer relativ großen Anzahl von Fachbereichen differieren die Fachbereichsmittelwerte um mehr als 0,5 Skalenpunkte. Die Ursachen für die starken Abweichungen in der Mathematik wurden oben bereits diskutiert, hinzu kommt, dass in der SPIEGEL Studie nicht nur Diplommathematiker sondern auch Lehramtsstudierende befragt wurden.

**Fazit: Der Vergleich der beiden unabhängig durchgeführten und methodisch abweichenden Ranking Studien zeigt, dass Erhebungen mit vergleichbarer Fragestellung und einer zumindest ähnliche definierten Grundgesamtheit zu weitgehend übereinstimmenden Ergebnissen kommen.**

**Abb. 6:** Vergleich der Studierendenbefragungen von CHE und SPIEGEL  
Gesamtbeurteilung der Studiensituation  
Mittelwerte und Mittelwertdifferenzen von Fachbereichen



#### 4. Einwand: Die Stichproben sind zu klein oder verzerrt

Die vergleichsweise kleine absolute Anzahl der Befragten Studierenden gibt immer wieder Anlass über zu kleine oder verzerrte Stichproben und deren Auswirkungen zu spekulieren (vgl. Ott 1999). Das Design von Hochschulrankings hat es nicht mit einer, sondern mit sehr vielen Grundgesamtheiten (jeder Fachbereich bildet eine Grundgesamtheit) zu tun. Die Gesamtzahl der Befragten liegt daher regelmäßig weit über der Stichprobengröße, die etwa für repräsentative Bevölkerungsbefragungen üblich ist. Allein für die Fächer Elektrotechnik und Maschinenbau/Verfahrenstechnik wurden z.B. in der CHE Befragung rund 65.000 Studierenden angeschrieben. Die Grundgesamtheit besteht allerdings nur aus den Studierenden im Hauptstudium. Deren Zahl liegt z.B. im Fach Maschinenbau an mehr als der Hälfte der Standorte unter 300 Studierenden, an mehr als 20% der Standorte sogar unter 100 Studierenden. An Standorten mit

weniger als 300 Studierenden im Hauptstudium wurden für den CHE Studienführer Vollerhebungen durchgeführt, an den übrigen Stichproben von 300 Studierenden gezogen. Das sind Stichprobengrößen, die die Standards der Umfrageforschung bei weitem übererfüllen.

Bis 50 Studierende	13,0 %
51 bis 100 Studierende	8,9 %
101 bis 200 Studierende	18,7 %
201 bis 300 Studierende	17,9 %
Über 300 Studierende	41,5 %

**Tab. 5: Fachbereiche nach Zahl der Studierenden zwischen dem 5. und 14. Fachsemester im Fach Maschinenbau, Universitäten und Fachhochschulen 1998. Quelle: Statistisches Bundesamt**

Ein Problem stellt allerdings der niedrige Rücklauf in den Studierendenbefragung dar (Studierenden je nach Fach 20% bis 25% Rücklauf; Professoren 40% bis 50%). Postalische Befragungen weisen grundsätzlich niedrigere Ausschöpfungsraten auf als persönlich-mündliche Befragungen. Empirisch variieren die Ausschöpfungsquoten postalischer Befragung zwischen 10% und 90%, und man geht von durchschnittlichen Werten in der Größenordnung von 47% aus (Porst 1999). Daran gemessen ist der Rücklauf in der Professorenbefragung befriedigend, in den Studierendenbefragungen jedoch nicht. Allerdings folgt aus einer niedrigen Ausschöpfungsraten keineswegs zwangsläufig auch eine Verzerrung der Stichprobe. Empirisch lässt sich zeigen, dass der sog. „Nonresponse-Bias“ keineswegs mit der Höhe der Ausschöpfungsquote abnimmt (Koch 1998). „Mehr“ bedeutet daher keineswegs auch „besser“. Wie der Vergleich mit der SPIEGEL Befragung (s.o.) gezeigt hat, ergeben sich zunächst keine Hinweise darauf, dass durch systematische Selbstselektion nur ein bestimmter Studierendentypus an den Befragungen teilgenommen hat. Die Behauptung, dass Stichproben allein aufgrund der niedrigen Ausschöpfungsraten verzerrt seien, ist eine Vermutung ohne Beleg. Replikationsstudien und der Vergleich unterschiedlicher Stichprobenziehungsverfahren sprechen bisher nicht für eine Verzerrung.

**Fazit: Der Stichprobenumfang des CHE Studienführers ist mehr als ausreichend. Die niedrigen Ausschöpfungsraten sind unbefriedigend, wenn auch nicht ungewöhnlich. Anhaltspunkte für eine systematische Verzerrung liegen bisher nicht vor.**

### **5. Einwand: Man braucht gar kein Ranking, die Größe der Hochschule bestimmt die Urteile der Studierenden**

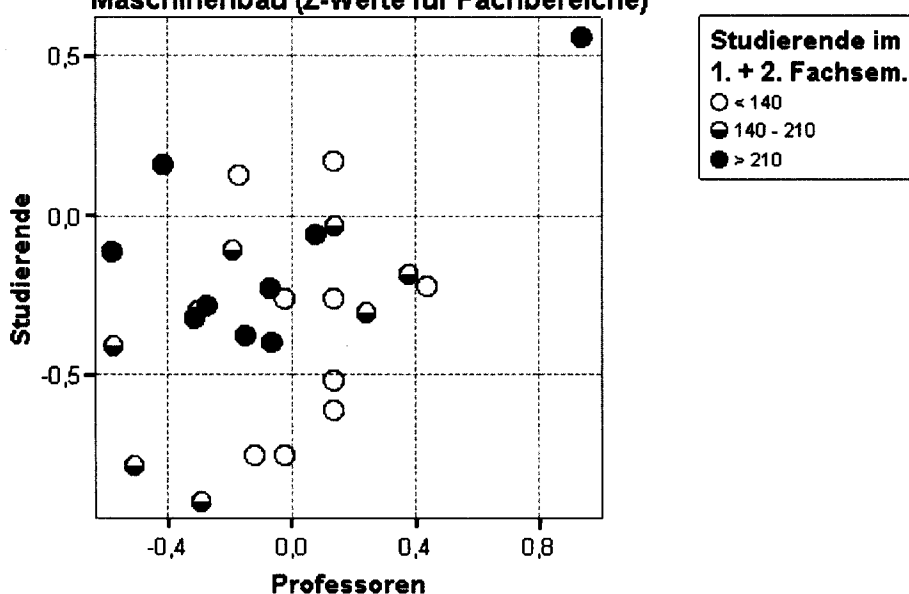
Hochschulrankings auf der Basis von Studierendenurteilen kommen immer wieder zu dem Ergebnis, dass kleinere Fachbereiche von den Studierenden tendenziell besser bewertet werden als große. Zur Erklärung wird darauf verwiesen, dass den Studierenden an den großen Fakultäten die Nestwärme fehle und sie allenfalls die Rahmenbedingungen, die nur indirekt etwas mit Qualität zu tun haben, beurteilen würden (vgl. Meinefeld 2000).

Betrachtet man einmal zwei Studiengänge genauer, einen stark nachgefragten mit hohen Studierendenzahlen pro Fakultät (Rechtswissenschaft) und einen weniger stark nachgefragten mit

vergleichsweise niedrigen Studierendenzahlen (Maschinenbau), dann lässt sich zeigen, dass solche Erklärungen wenig empirischen Rückhalt haben: Zunächst einmal zeigen die Graphiken 7 und 8, dass die Beziehung zwischen Größe und Bewertung keineswegs in allen Fächern gilt, sondern vor allen Dingen in den großen, überlaufenen Studienfächern. Im Maschinenbau sind derartige Beziehungen kaum zu erkennen. Sodann deuten die gleichgerichteten Urteile von Professoren und Studierenden darauf hin, dass auch die Professoren an großen Universitäten die Lehrsituation schlechter beurteilen. Dies kann nicht auf fehlende Nestwärme zurückzuführen sein. Schließlich zeigt sich sowohl im Maschinenbau wie auch in der Rechtswissenschaft, dass es großen Hochschulen durchaus gelingen kann, positive Urteile von Studierenden und Professoren zu erhalten. Die unterschiedlichen Bewertungen der großen Hochschulen zeigen zudem, dass auch unter widrigen Bedingungen sich sehr unterschiedliche Ergebnisse produzieren lassen. Schlussendlich spricht etwa in der Rechtswissenschaft auch nichts dafür, dass große Fakultäten eine bessere, von den Studierenden nicht gewürdigte, Lehrleistung erbringen; die Noten im Staatsexamen unterscheiden sich nicht signifikant von denen an kleineren Fachbereichen.

Abb. 7:

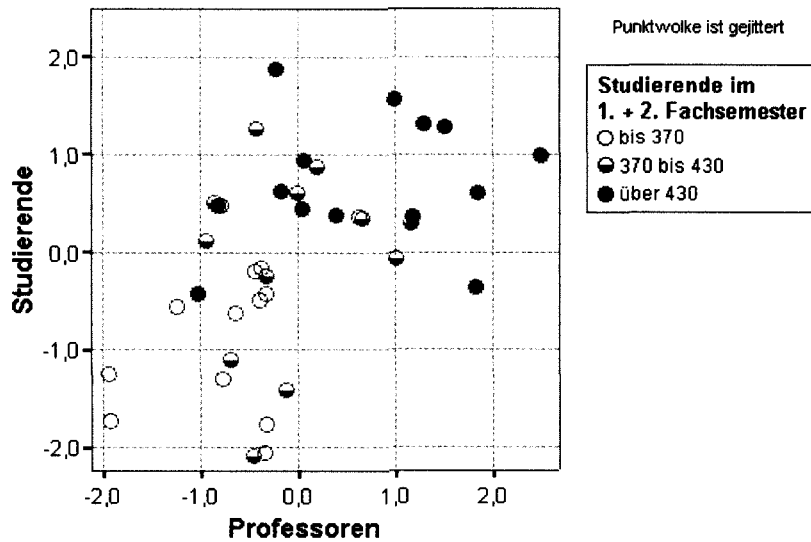
### Gesamturteile zur Lehre von Studenten und Professoren im Maschinenbau (Z-Werte für Fachbereiche)



CHE Studienführer 2000

**Fazit:** Die Größe von Fachbereichen determiniert keineswegs die Urteile von Studierenden und Professoren, vielmehr scheint sich nur in den stark frequentierten Studienfächern die Größe der Fachbereiche auf die Lehrqualität auszuwirken, und selbst dann sind große Fachbereiche durchaus in der Lage, Lehr- und Betreuungsangebote zu entwickeln, die von Studierenden und Professoren positiv beurteilt werden.

**Gesamturteile zur Lehre von Studenten und Professoren in der Rechtswissenschaft (standardisierte Z-Werte für Fachbereiche)**



**Abb. 8:** Rechtswissenschaft

### 6. Einwand: Mittelwert- und Indexbildungen nivellieren die Urteile

Die Fülle von Informationen, die in Rankingstudien typischerweise erhoben wird, muss für den Nutzer in geeigneter Form verdichtet werden. Das gilt sowohl für die Zahl der Variablen als auch für die Charakterisierung der studentischen Urteile an einem Fachbereich.

Erstere Informationskomprimierung wird üblicherweise durch die Entwicklung von Indices oder Skalen bewerkstelligt, die möglichst eindimensional die Eigenschaften des Bewertungsobjektes wiedergeben sollen, letztere über die Berechnung von Mittelwerten und geeigneten Angaben zur Homogenität bzw. Streuung der gemittelten Werte. Indices, wie sie auch im CHE Studienführer benutzt werden, durchlaufen also eine zweimalige Mittelwertbildung. Die Folge ist, dass nicht mehr erkennbar ist, ob mittlere Urteile oder sehr heterogene Urteile dem Mittelwert zugrunde liegen. Zwar informieren die berechneten Konfidenzintervalle darüber, wie einheitlich die Indexwerte ausgefallen sind (s.u.), aber nicht mehr darüber, ob die Befragten typische Urteilsprofile aufweisen (vgl. auch Kromrey 2000).

Es ist deshalb sicherlich wünschenswert nach Möglichkeiten einer informationshaltigeren Aufbereitung der Daten zu suchen. Kromrey (vgl. den Beitrag in diesem Band) schlägt dazu vor, die Urteilenden aufgrund von clusteranalytisch gewonnenen Urteilstypologien in verschiedene Beurteilungscluster einzuordnen und eine Bewertung der Fachbereiche auf die quantitativ dominanten Cluster zu stützen. Für eine Routineberichterstattung ist dies Verfahren leider nicht tauglich. Clusteranalysen reagieren sehr empfindlich auf die gewählten Fusionierungsverfahren, Ähnlichkeitsmaße und die Behandlung fehlender Werte. Hinzu kommt, dass ein solches Verfahren – je nach Fach – bis zu 60 Einzelurteile der Studierenden zu einem Profil verarbeiten müsste. Unter solchen Bedingungen entstehen entweder beliebige oder triviale Ergebnisse, denn einfache Antwortmuster wie „alles gut“ oder „alles schlecht“ gibt es so gut wie nicht. Ihre Existenz

würde auch die Frage nahe legen, ob hier nicht ein „response set“ wirksam ist, also eine stereotype Bewertung unabhängig von den gestellten Fragen. In aller Regel antworten die Studierenden sehr differenziert. Selbst wenn man die Kriterien für „alles gut“ sehr weit fasst, bleibt diese Gruppe in der Minderheit. Betrachtet man beispielsweise nur einmal die Fragenkomplexe zur Lehre, Studienorganisation, Raumausstattung und Bibliothek, dann antworten von den ca. 23.000 Studierenden der CHE Befragung 1999 nur knapp 20% so, dass mindestens zwei Drittel der Items mit den Skalenwerten 1 und 2 benotet werden, während ca. 80% gute und schlechte Noten so vergeben, dass keine durchgängige Präferenz für eine Note erkennbar ist. Angesichts der oben genannten Fallzahlen produzieren Clusteranalysen schließlich für kleine Fachbereiche ein zusätzliches Problem: Die Fallzahl in den einzelnen Clustern kann sehr niedrig werden, eine Interpretation ist dann kaum mehr möglich.

Am Beispiel des Index Lehre aus der Studierendenbefragung im Maschinenbau lassen sich Nutzen und Problematik clusteranalytischer Gruppierung demonstrieren<sup>6</sup>. Tabelle 6 zeigt eine Vier-Clusterlösung für die acht Einzelitems, die im Index Lehre verrechnet werden<sup>7</sup>. Cluster 4 weist eine sehr positive Urteilstendenz auf, das Gegenstück (Cluster 2) zeigt hingegen eine sehr kritische Beurteilung der acht Items. Cluster 1 und Cluster 3 zeigen die erwartete Differenzierung in den mittleren Urteilen. Im Cluster 1 wird etwas schlechter als im Durchschnitt geurteilt, wobei insbesondere die Praxisdimension (Lehrveranstaltungen durch Praktiker, Praxisbezug, Interdisziplinarität) sehr kritisch beurteilt wird. Im Cluster 3 hingegen wird etwas besser als im Durchschnitt geurteilt (mit Ausnahme der Didaktik und des Angebots an Projektseminaren). In der Zeile „Index“ sind zum Vergleich die Mittelwerte des Index Lehre angegeben; sie fassen die Urteilstendenz in diesen Clustern durchaus treffend zusammen. Insgesamt stellen die Cluster also eine Mischung aus Niveaushiftung der Urteile und schwach ausgeprägten Urteilsprofilen dar.

	Maschinenbau Cluster			
	Cluster 1 25% der Personen	Cluster 2 11,5% der Personen	Cluster 3 43 % der Personen	Cluster 4 21 % der Personen
	Mittelwert	Mittelwert	Mittelwert	Mittelwert
Lehrangebot Projektseminare etc.	2,87	4,17	2,55	1,63
Lehrangebot inhaltliche Breite	2,07	2,66	1,89	1,51
Lehrangebot Forschung	3,19	3,82	2,49	1,81
Lehrangebot Praxisbezug	3,61	4,71	2,64	1,77
Lehrangebot internat. Ausrichtung	3,30	4,20	2,89	1,83
Lehrangebot Interdisziplinarität	3,14	4,25	2,44	1,76
Lehrangebot Vermittlung Lehrstoff	3,47	4,48	2,71	2,07
Lehrangebot Praktiker-Veranstalt.	4,07	4,76	2,79	1,75
INDEX	3,22	4,13	2,55	1,77

Tab. 6 ( in der Zeile Index sind die Mittelwerte für den berechneten Index Lehre angegeben)

<sup>6</sup> Das Beispiel von Kromrey (in diesem Band) kombiniert den Index Lehre mit dem Index Studienorganisation, so dass die unterschiedlichen Ranggruppen nicht vergleichbar sind. Der Index Lehre, der im CHE Studienführer benutzt wurde, besteht aus acht einzelnen Items. Für die folgenden Analysen wurden nur Fälle berücksichtigt, die auf allen acht Items valide Werte aufweisen.

<sup>7</sup> Vgl. dazu den Methodenbericht zum CHE Studienführer: CHE Arbeitspapier Nr. 22, 2000

Betrachtet man nun, wie sich die Befragten an den einzelnen Hochschulen auf die Cluster verteilen (vgl. Tabelle 7), dann zeigt sich, dass erwartungsgemäß an den gut beurteilten Hochschulen (kleiner Indexmittelwert) das Cluster 4 überwiegt und umgekehrt an den schlecht beurteilten die Cluster 2 und 1. Allerdings ist Cluster 2 an keiner Hochschule stärker besetzt als die übrigen Cluster; das würde bedeuten, dass eine Schlussgruppe gar nicht existiert. Eine solche Interpretation wird den Daten aber erkennbar nicht gerecht, denn an den schlecht bewerteten Hochschulen (z.B. Hannover, Berlin, Wuppertal) findet sich eine deutliche Mehrheit in den kritisch urteilenden Clustern 1 und 2, während typische Mittelfeld Universitäten (z.B. Duisburg, Dresden) eine deutliche Mehrheit im Cluster 3 aufweisen.

**Tab 7: Maschinenbau - Index Lehre**  
Prozentuale Besetzung der Cluster, Indexmittelwerte und Fallzahlen

			% in	% in	% in	% in	N <sup>a</sup>
			Cluster 1	Cluster 2	Cluster 3	Cluster 4	
Index	2,13	TU Clausthal	8,3	2,3	39,4	50,0	132
Lehre	2,14	Bergakad. Freiberg TU	11,4	,0	38,6	50,0	44
MW	2,20	U Kiel	20,0	,0	40,0	40,0	5
	2,30	TU Ilmenau	25,9	,0	33,3	40,7	27
	2,36	U Bremen	20,0	2,9	45,7	31,4	35
	2,42	U Rostock	17,4	8,7	34,8	39,1	23
	2,46	TU Chemnitz	17,6	2,9	52,9	26,5	34
	2,60	U Saarbrücken	23,1	3,8	50,0	23,1	26
	2,62	UGH Siegen	22,7	6,8	43,2	27,3	44
	2,62	TU Hamburg-Harburg	22,7	7,3	50,9	19,1	110
	2,63	RWTH Aachen	20,4	14,3	40,8	24,5	49
	2,63	BTU Cottbus	25,6	6,1	46,3	22,0	82
	2,66	UGH Paderborn	11,5	15,4	42,3	30,8	26
	2,69	TU Dresden	19,8	7,2	61,3	11,7	111
	2,70	UGH Duisburg	22,0	10,0	50,0	18,0	50
	2,73	U Kaiserslautern	25,7	5,7	48,6	20,0	35
	2,77	TU Darmstadt	33,3	10,4	37,5	18,8	48
	2,82	U Dortmund	25,6	14,1	34,6	25,6	78
	2,84	UGH Kassel	19,4	13,9	52,8	13,9	36
	2,86	U Erlangen-Nürnberg	28,9	10,0	44,4	16,7	90
	2,87	U Stuttgart	26,5	17,1	46,2	10,3	117
	2,90	TU Braunschweig	25,9	19,0	36,2	19,0	58
	2,92	U Karlsruhe (TH)	28,3	18,9	37,7	15,1	106
	2,93	U Bochum	30,0	20,0	35,0	15,0	20
	2,97	UGH Essen	29,0	16,1	45,2	9,7	31
	3,01	TU München	41,5	11,3	37,7	9,4	53
	3,18	U Hannover	26,9	26,9	34,6	11,5	26
	3,27	TU Berlin	41,6	23,0	31,9	3,5	113
	3,36	UGH Wuppertal	31,3	25,0	40,6	3,1	32

<sup>a</sup>. Nur Fälle, die auf allen 8 Variablen, die für die Indexbildung benutzt wurden, valide Werte aufweisen.

Für einen Vergleich liegt es daher nahe, die Cluster 1 und 2 zu einer Schlussgruppe zusammenzufassen. Gruppiert man nun die Hochschulen danach, welches Cluster den größten Anteil der Befragten auf sich vereinigen kann, dann ergibt sich eine Spitzengruppe (Cluster 4 dominiert), eine Mittelgruppe (Cluster 3 dominiert) und eine Schlussgruppe (die Summe von Cluster 1 und 2 dominiert). Diese Gruppeneinteilung lässt sich mit der im CHE Studienführer angewandten Ranggruppenbildung auf der Grundlage von Konfidenzintervallen um den Mittelwert (s.u.) vergleichen.

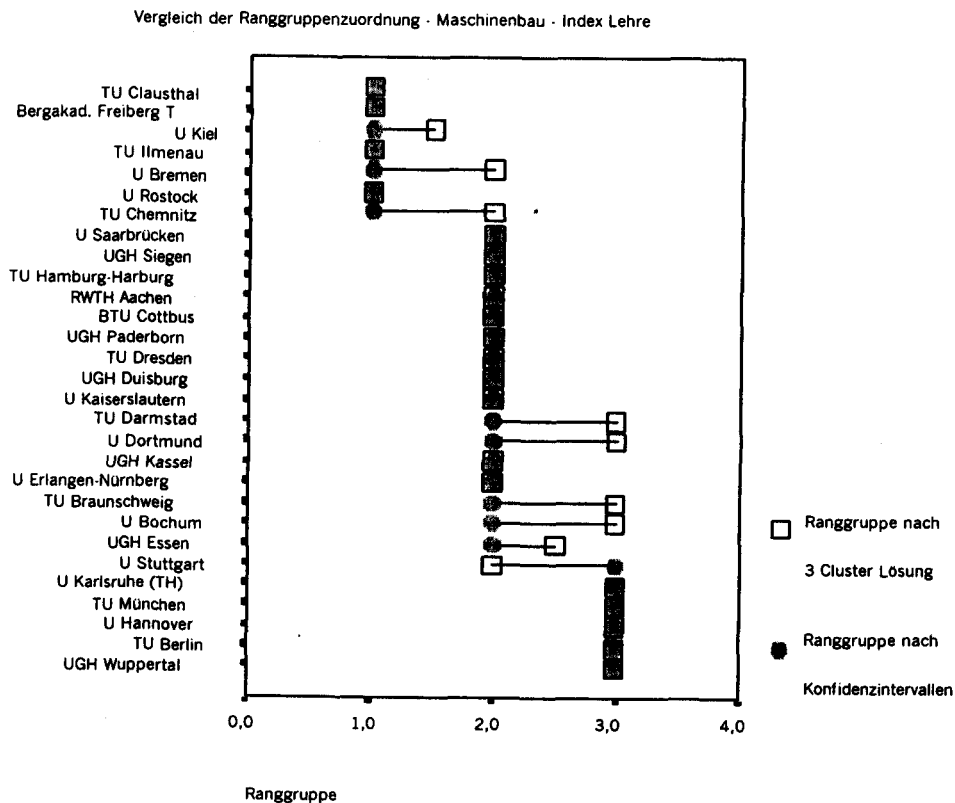


Abb. 9: Ranggruppen nach Clustermodell und nach Konfidenzintervallen

Das Ergebnis ist in Abbildung 9 wiedergeben. Erkennbar ist zunächst, dass die clusterbasierte Gruppenzuordnung keine grundsätzlich anderen Ergebnisse hervorbringt, als die auf Mittelwerten und Konfidenzintervallen basierende Gruppeneinordnung. Wechsel finden nur zwischen der Mittelgruppe und den Extremgruppen statt, nicht aber zwischen den Extremgruppen. Insgesamt überwiegt die Zahl gleich klassifizierter Hochschulen deutlich die Zahl der Abweichungen. Die Abweichungen treten bevorzugt an den Gruppengrenzen auf, also dort, wo eine Abgrenzung ohnehin schwerfällt. Weiterhin ist erkennbar, dass von den neun Hochschulen, die die Ranggruppe wechseln, zwei nicht eindeutig verortet werden können (in Kiel und Essen weisen zwei Cluster die gleiche Besetzungshäufigkeit auf). Und schließlich ist zu berücksichtigen, dass die Hochschulen Essen, Bochum, Chemnitz, Bremen und Kiel<sup>8</sup> in dieser Reanalyse (vgl. die Ausschlusskriterien und Tab. 7) nur noch mit Stichprobengrößen von  $\leq 35$  Personen repräsentiert sind. Die clusterbasierte Gruppenzuordnung wird damit sehr instabil: Beispielsweise kann an der Universität Bochum eine geringfügige Veränderung der Urteile von nur zwei Personen dazu führen, dass ein anderes Cluster dominant wird. Auch an der Universität Stuttgart liegt das Kritikercluster (43,6%) und das mittlere Cluster (46,2%) extrem nah beieinander. Ähnliches gilt für Dortmund und Darmstadt.

<sup>8</sup> Die Universität Kiel bietet keinen Studiengang Maschinenbau an. Sie ist in diesem Sample nur enthalten, weil unter die Definition „Studienbereich Maschinenbau“, wie sie vom CHE festgelegt wurde, auch der Studiengang „Materialwissenschaft“ gehört, der in Kiel angeboten wird.

Ein Zugewinn an Informationen entsteht also nicht auf der Ebene von Ranggruppenzuordnungen, sondern nur dann, wenn man die Besetzung aller vier Cluster im Auge behält. Dann lässt sich etwa die überwiegend durchschnittlich beurteilte TU Dresden von den sehr heterogen beurteilten Hochschulen Braunschweig oder Darmstadt unterscheiden. Dieser Informationszugewinn bedeutet aber auch, dass aus einer Kennzahl (Index Lehre) vier Kennzahlen (Besetzungshäufigkeiten in den vier Clustern) werden. Dieser Weg der Darstellung bietet sich also für vertiefende Information an, allerdings stellt sich die Frage, ob dann nicht an Stelle der tendenziell instabilen und von vielen Konstruktionsentscheidungen abhängigen Clusteranalyse eine einfache Darstellung der Notenverteilung innerhalb eines Indexwertes nicht ebenso informativ, wenn nicht sogar informativer ist. Abbildung 10 zeigt eine solche Verteilung für einige der oben analysierten Daten. Vergleicht man die nahe beieinander im Mittelfeld platzierten Hochschulen Aachen und Dresden, zeigt sich sehr deutlich, dass Aachen einen weitaus höheren Anteil an „Einser-Noten“ auf den acht Items des Index Lehre verbuchen konnte als Dresden.

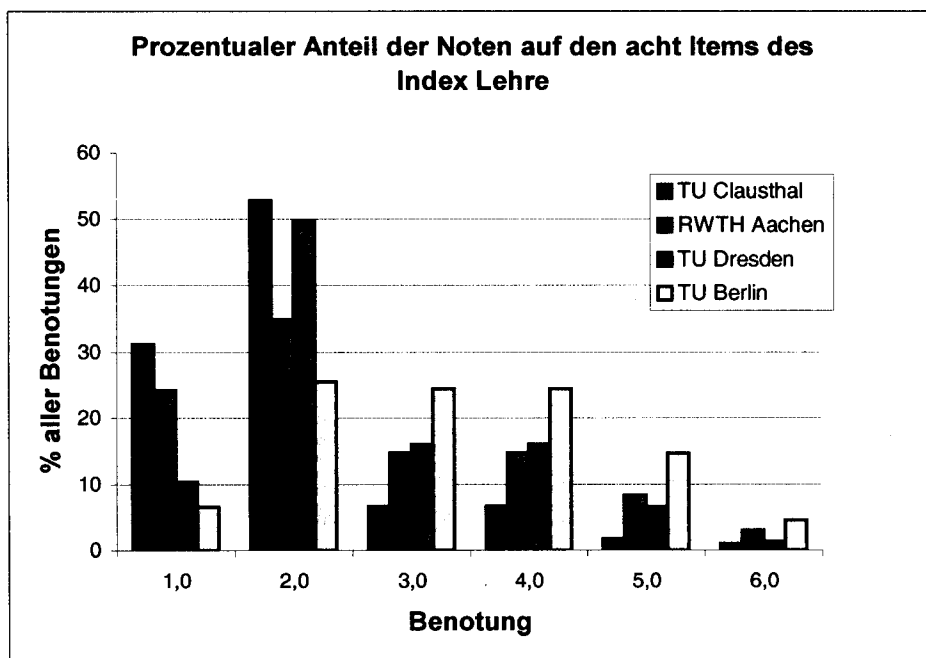


Abb. 10

Ein anderer Weg, die Informationen über die Urteilsstreuung besser zu nutzen, besteht darin, theoretisch plausible und empirische vorfindbare Differenzierungen für eine Subgruppenbildung zu nutzen (vgl. oben das Beispiel Mathematik). Man überlässt dann die Gruppenbildung nicht den Unwägbarkeiten der Clusteranalyse, sondern benutzt Grundorientierungen (z.B Praxis- versus Forschungsorientierung), um homogene Subpopulationen zu erzeugen.

**Fazit: Index- und Mittelwertbildungen stellen Informationsverdichtungen dar, die notwendigerweise auch zu Informationsreduktionen führen. Der Vergleich von clusteranalytischen Verfahren und varianzanalytisch basierten Verfahren zeigt aber, dass bei einer hohen Informationsverdichtung – wie sie für orientierende Überblicke notwendig ist – keine gravierenden Differenzen auftreten. Cluster- und Subgruppenanalysen stellen sehr wünschenswerte ergänzende und vertiefende Information zur Verfügung, einen Ersatz für**

**Mittelwertbildungen stellen sie jedoch aufgrund der Verfahrensabhängigkeit und Instabilität nicht dar.**

### **7. Einwand: Die Hochschulen unterscheiden sich kaum, Ranggruppen führen zu arbiträren Urteilen**

Das CHE verzichtet völlig auf die Vergabe von Rangplätzen, weil es unsinnig ist, den Fachbereichen Ränge zuzuweisen, die eine Genauigkeit suggerieren, welche durch stichprobenbasierte Befragungsdaten nicht einzulösen ist. Deshalb werden Konfidenzintervalle um die Fachbereichs-Mittelwerte der jeweiligen Urteile ermittelt, und auf dieser Grundlage wird zunächst eine Spitzen- und eine Schlussgruppe gebildet. Zwischen den Hochschulen dieser beiden Gruppen bestehen erhebliche und statistisch signifikante Urteilsdifferenzen. Dabei werden die Stichprobengröße und die Homogenität der Urteile am Fachbereich berücksichtigt. Alle Fachbereiche, an denen die Befragten sehr uneinheitlich geantwortet haben und/oder insgesamt ein "durchschnittliches" Urteil abgeben, fallen in eine mittlere Ranggruppe. Die dort platzierten Fachbereiche unterscheiden sich weder vom Durchschnitt des Faches noch von den benachbarten Fachbereichen in der Spitzen- oder Schlussgruppe statistisch signifikant. Die Größe des Konfidenzintervalls informiert zudem darüber, ob hier weitgehend mittlere Urteile abgegeben wurden oder ob sehr heterogene Urteile zu einem mittleren Mittelwert geführt haben. Es ist der Sinn des Ranggruppenmodells, die Interpretation auf die Hochschulen in der Spitzen- und Schlussgruppe zu begrenzen. Dabei handelt es sich nicht um eine kosmetische Beigabe zu den klassischen Ranglisten, vielmehr bilden die Ranggruppen im Studienführer des CHE die Basis für eine nutzerspezifische Zusammenstellung von Fachbereichen, entweder mit Hilfe einer CD Rom, die dem Studienführer des CHE beigelegt ist oder über ein „persönliches Ranking“ auf der Internetseite des Studienführers (<http://www.stern.de/servlet/CHE2>).

Auch Statistikern fällt es jedoch schwer, dieser Interpretation zu folgen. So fragt z.B. Jensen (vgl. den Beitrag in diesem Band), warum in der Elektrotechnik die Universitäten Karlsruhe und Siegen mehr mit der Universität Chemnitz gemeinsam haben sollten, als mit Magdeburg oder Kiel. Ein Blick auf die Konfidenzintervalle beantwortet diese Frage sehr schnell: Eine solche Behauptung wurde nie aufgestellt. Die Konfidenzintervalle zeigen sehr klar, dass Karlsruhe und Siegen **weniger** mit der Universität Chemnitz gemeinsam haben als mit Magdeburg oder Kiel. Ausgesagt ist lediglich, dass Karlsruhe und Siegen besser bewertet sind als die Schlussgruppe (z.B. Hannover, Wuppertal), was sich eben von Kiel nicht behaupten lässt.

Diese Befunde sind darüber hinaus weitgehend verfahrensunabhängig. Berechnet man anstelle der Konfidenzintervalle von Verteilungsannahmen unabhängige Bootstrap-Rangkonfidenzintervalle (vgl. Jensen in diesem Band), entstehen extrem ähnliche Ergebnisse (so sticht beispielsweise hier wie dort die BTU Cottbus mit einem sehr breiten Konfidenzintervall heraus).

## Elektrotechnik - Universitäten

### 95% Konfidenzintervall um den Mittelwert des stud. Gesamturteils

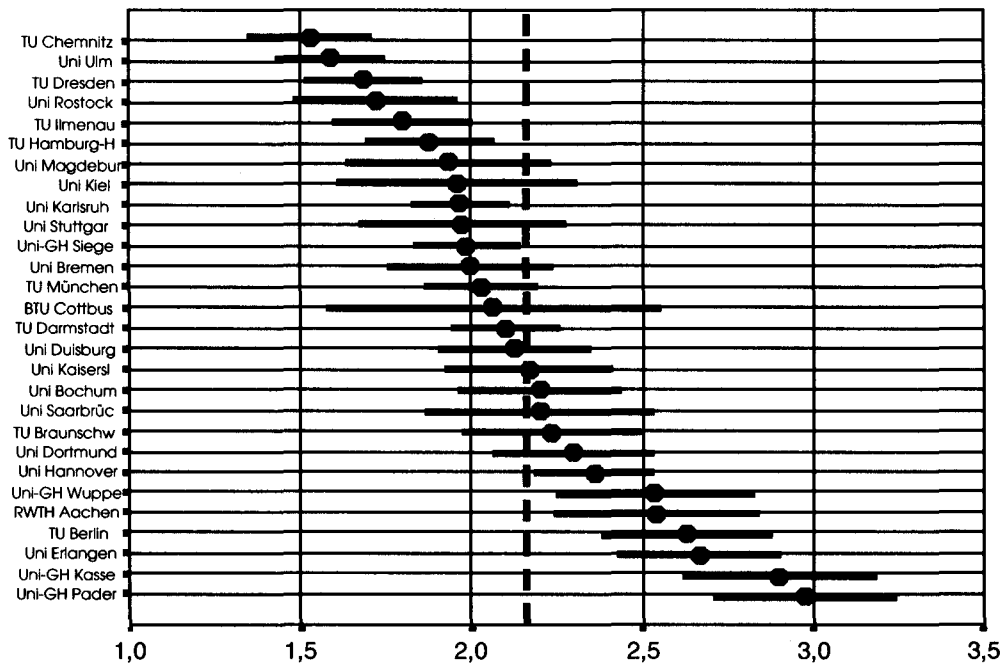


Abb. 11

Diese Ergebnisse sind auch durchaus plausibel, denn anders als in der BWL oder Rechtswissenschaft liegen in den Ingenieurwissenschaften vergleichsweise homogene Verhältnisse vor. Ein Ranking hat nun nicht die Aufgabe, künstliche Differenzen aufzubauen, sondern durchaus auch die Aufgabe, homogene Verhältnisse abzubilden. Ranggruppen entstehen nur dann, wenn tatsächlich bedeutsame Urteilsunterschiede vorhanden sind. Ihre Größe hängt davon ab, wie stark die Beurteilungen an den Fachbereichen voneinander abweichen. Unter Bedingungen völliger Homogenität würde sich nach dem Ranggruppenmodell nur eine einzige große Mittelgruppe ergeben. Natürlich lässt sich mit einem solchen Modell kein "Sieger" ermitteln, sondern nur eine Gruppe von Hochschulen, die signifikant besser als der Durchschnitt im Fach beurteilt wird, und eine Gruppe, die signifikant schlechter beurteilt wird. Eine derartige Begrenzung passt aber nicht gut zum verbreiteten Bedürfnis Hochschulen auf ein Siegetreppchen zu stellen, das aufgrund von Rangplätzen besetzt wird (vgl. Meinefeld): Wenn die publizierten Hitlisten – wie im Falle des CHE-Studienführers – keine Rangplätze enthalten, dann werden sie einfach durch Auszählung im Nachhinein ermittelt, um dann festzustellen, dass die Juristen der Universität Potsdam (Mittelgruppe) aufgrund der studentischen Beurteilungen zwischen dem 9. und 28. Rangplatz positioniert sein könnten. Ein solcher Befund ist nun bestenfalls eine Tautologie, denn nichts anderes besagt die Einstufung in die mittlere Ranggruppe (vgl. ebd.).

Was jedoch unbefriedigend an jeder Gruppeneinteilung bleibt, das sind die überkonturierten Grenzziehungen zwischen den Gruppen. Dieses Problem entsteht bei jeder Gruppenbildung, unabhängig vom gewählten Verfahren: So entscheiden z.B. im von Kromrey vorgeschlagenen Clustermodell extrem wenige Personen darüber, ob ein Cluster die 40% oder 50% Marke übersteigt. Zur nachfolgenden Hochschule besteht dann i.d.R. natürlich kein signifikanter Unter-

schied. Ranggruppen stellen lediglich eine Dichtomisierung zwischen sehr gut und sehr schlecht bewertet her. Dabei bleiben die Grenzen zur Mittelgruppe notwendigerweise unscharf. Um einen Vergleich zwischen einzelnen Hochschulen zu ermöglichen, wäre es daher sinnvoll neben den Ranggruppen auch die Konfidenzintervalle für die einzelnen Mittelwerte zu präsentieren (vgl. z.B. Hornbostel 1998, 1999b; Müller-Böling und Hornbostel 2000). Dann würde sehr schnell deutlich, dass Universitäten wie Kiel oder Magdeburg aufgrund der heterogenen Urteile und der entsprechend breiten Konfidenzintervalle in der Mittelgruppe positioniert wurden.

**Fazit: Rankingkritiker scheinen sich nur schwer damit abfinden zu können, dass zwischen sehr vielen Hochschulen keine signifikanten Bewertungsunterschiede bestehen. Die üblicherweise berechneten Konfidenzintervalle geben jedoch – auch im Vergleich mit anderen Gütebeurteilungen – sehr zuverlässig an, wo Unterschiede behauptet werden können und wo nicht. Wünschenswert wäre es, wenn diese Information – ergänzend zu den Ranggruppen - auch in den publizierten Rankings zu finden wäre.**

### **8. Einwand: Zu viel subjektive Urteile, zu wenig harte Fakten.**

In der Tat enthält der Studienführer des CHE überwiegend subjektive Urteile. Zwar werden eine ganze Reihe von statistischen Rahmendaten ausgewiesen, ebenso die Fachstudiendauer, aber andere Kernbereiche werden nicht mit Fakten versehen. Das hat verschiedene Gründe: Manche Informationen wie etwa die Teilnehmerzahlen an Veranstaltungen oder Universitäts-Eingangstests, wie in den USA üblich, existieren einfach nicht, andere Informationen wie etwa Abbruchquoten sind aufgrund gesetzlicher Regelung (Einstellung der Individualstatistik) nicht verfügbar, wieder andere existieren nur rudimentär, wenngleich sie sehr wünschenswert wären wie z.B. Informationen über den beruflichen Werdegang der Absolventen.

Es gibt auch eine Reihe von Informationen, die prinzipiell zugänglich sind, aber nicht zu interpretationsfähigen Indikatoren verarbeitet werden können. Dazu gehört beispielsweise, dass es bisher nicht gelungen ist, eine Betreuungsrelation, die auch die Dienstleistungsverflechtungen angemessen berücksichtigt, einheitlich zu definieren und zu erfassen (also das tatsächliche numerische Verhältnis von Professoren und zu betreuenden Studierenden, nicht die normativ fixierten Curricularnormwerte).

Strukturdaten sind ein Beispiel dafür, dass es nicht sinnvoll ist, von vornherein die Interpretation solcher Kennzahlen auf einer „gut-schlecht“ Dimension festzulegen, sondern dem Informationsnutzer Interpretationsspielräume zu überlassen. Ein hoher Frauenanteil unter den Studierenden ist zunächst weder gut noch schlecht, er kann aber eine solche Konnotation bekommen, wenn etwa eine Schülerin sich für ein männerdominiertes Fach interessiert und bei der Standortentscheidung den Frauenanteil berücksichtigt. Ähnliches gilt für die Größe von Fachbereichen und Hochschulen, die Studiendauer etc.

Werden Kennzahlen jedoch für evaluative Zwecke eingesetzt, ist eine solche Beliebigkeit äußerst unerwünscht. Und dies ist denn auch die Achillesferse „harter Daten“. Beispielsweise wurde versucht für den CHE Studienführer die Qualität der Bibliothek zusätzlich zu den subjektiven Einschätzungen durch harte Fakten aus der Bibliotheksstatistik zu ergänzen. Nur leider ist es nur begrenzt möglich – abgesehen davon, dass Institutsbibliotheken gar nicht erfasst werden – Abonnementszahlen, laufende Ausgaben oder Monographiebestände als Qualitätsindika-

toren zu interpretieren. Harte Fakten sind nicht per se besser als subjektive Einschätzungen, wenn es darum geht die Eignung eines Angebotes für die Bedürfnisse der Nutzer zu beurteilen.

Was allerdings an „harten Fakten“ verfügbar und interpretierbar, steht auf der CD Rom des Studienführers auch zur Verfügung.

### **9. Einwand: Rankings reduzieren das komplexe Leistungsspektrum von Hochschulen auf studentische Beurteilungen.**

Rankingkritiker stellen häufig auf die Mehrdimensionalität hochschulischer Leistungen ab, die mit studentischen Befragungen allein nicht eingefangen werden kann. Das ist zweifellos richtig, erstaunlicherweise wird dabei allerdings die wichtigste Leistungsdimension neben der Lehre, nämlich die Forschung, kaum thematisiert. Auch US-amerikanische Rankings berücksichtigen diese Dimension in der Regel nicht. Dass Forschung und Lehre ein Kuppelprodukt bilden, heißt weder, dass sich automatisch Humdoldtsche Ideale einer die Lehre befruchtenden Forschung einstellen, noch umgekehrt, dass sich Forschung und Lehre in einer Art antagonistischer Zuspitzung gegenseitig behindern (vgl. Schimank 1995, Güdler/Sack1996). Vielmehr existiert eine bunte Mischung aus allen denkbaren Kombinationen.

Forschungsleistungen stellen nicht nur die wichtigste Quelle für das Renommee einer Hochschule dar, sie sind auch für jene Studierenden mit einer starken Forschungsorientierung, ebenso wie für Doktoranden, Hochschulwechsler und last not least für die an Forschung interessierten Dritten eine wichtige Information (vgl. Weingart 1995, Hornbostel 1999c).

Forschungsleistungen lassen sich insofern einfacher als Lehrleistungen erfassen, als diese i.d.R. öffentlich zugänglich sind (publiziert) und eine überlokale community laufend über die Qualität – zumindest eines Teils - dieser Forschungsergebnisse, Urteile fällt (Manuskriptbegutachtungen Drittmittelbewilligungen, Tagungseinladungen, Auszeichnungen, Zitierungen etc.). Allerdings gilt für Forschungsindikatoren ebenso wie für die Lehre, dass sie a) fachspezifisch justiert werden müssen (bzw. nur fachspezifisch einsetzbar sind), b) geeignete (meist umstrittene) Bezugsgrößen definiert werden müssen (Personal), c) fast immer subdisziplinäre Differenzierungen vorliegen, die ähnlich wie bei den Lehrindikatoren die Frage nach geeigneten Maßzahlen bzw. Differenzierungen in der Darstellung aufwerfen.

Das zeigt sich beispielsweise deutlich bei einer Analyse der Drittmiteleinwerbungen. Sie sind vor allen Dingen deshalb interessant, weil sie anders als etwa bibliometrische Indikatoren, die nur retrospektiv berichten, sehr zeitnahe oder sogar prospektive Informationen liefern. Bei diesem Indikator ist allerdings strittig, ob man ihn als Input oder Output Messung interpretieren kann. Für erstere Deutung spricht, dass die Miteleinwerbungen noch nichts über den Ertrag eines Projektes aussagen. Für letztere spricht, dass erstens der Drittmittelbewilligung i.d.R. ein aufwändiger Begutachtungsprozess zugrunde liegt, zweitens mit dem Drittmittelantrag häufig bereits ein sehr voraussetzungsvolles Zwischenprodukt vorliegt. Ein Indiz dafür, dass die Begutachtung von Drittmittelanträgen durchaus prognostische Validität besitzt, ergibt sich, wenn man einmal den Publikationsoutput daraufhin untersucht, ob die aus drittmittelgeförderten Projekten hervorgegangenen Publikationen eine andere fachliche Resonanz erzeugen als die übrigen Publikationen (vgl. Hornbostel 1997).

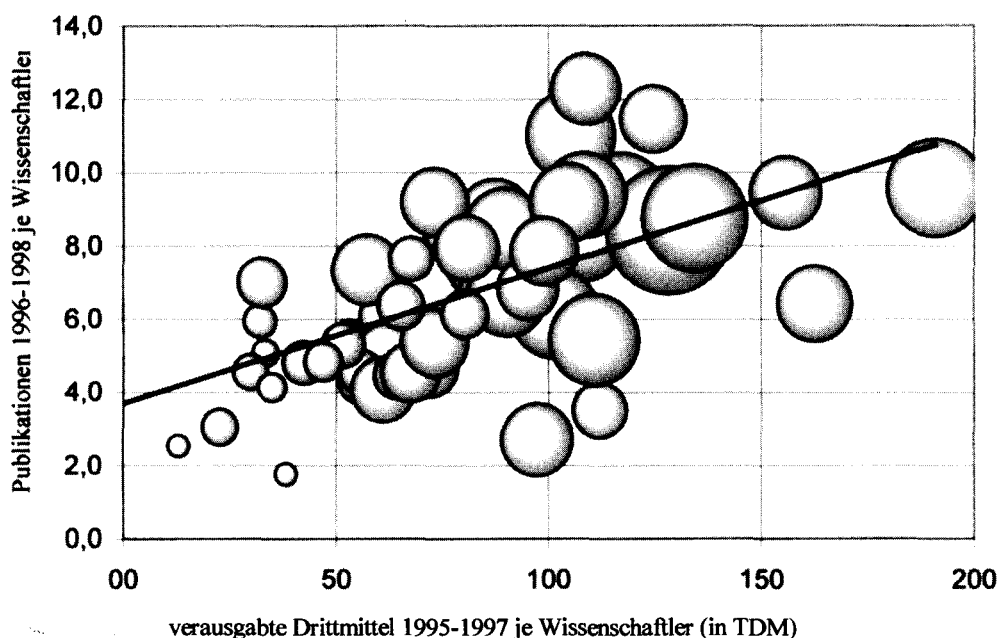
In welchem Umfang aus Drittmittelprojekten Publikationen in hochrangigen Zeitschriften resultieren, hängt allerdings auch davon ab, wie die Publikationsgepflogenheiten im jeweiligen Spezialgebiet beschaffen sind, welche Drittmittelanteile für apparative Zwecke und sonstige

Infrastruktur benötigt werden, dem time-lag zwischen Forschungsprojekt und Publikationszeitraum und schließlich natürlich vom Erfolg des Forschungsprojektes. Abb. 12 zeigt für die Physik die Anzahl der Beiträge in Zeitschriften (pro Professor am Fachbereich), die im Science Citation Index ausgewertet werden und der Höhe der verausgabten Drittmittel. Deutlich erkennbar ist die Beziehung zwischen Drittmittelvolumen und Publikationen, aber auch der Umstand, dass im mittleren Bereich pro „Drittmittelmark“ sehr unterschiedlich hohe Publikationszahlen realisiert werden.

Gewichtiger als subdisziplinäre Differenzen sind die Disziplinunterschiede. Abb. 13 zeigt die Antragsaktivität in den einzelnen Disziplinen. Die Daten basieren auf einer Vollerhebung unter den Professoren der jeweiligen Fachgebiete im Rahmen des Studienführers des CHE (vgl. Hornbostel 2001a). Von den rechtswissenschaftlichen Professoren haben mehr als 60% in den letzten drei Jahren keine Drittmittelanträge gestellt. Dieser Anteil liegt bei den Maschinenbauern unter 10%, dafür haben aber fast 40% der Maschinenbauprofessoren mehr als 10 Drittmittelanträge innerhalb von drei Jahren gestellt. Dieser unterschiedlichen Antragsintensität korrespondieren natürlich entsprechende Unterschiede im Gesamtvolumen der eingeworbenen Drittmittel.

Von den befragten Professoren aus den Bereichen Maschinenbau/Verfahrenstechnik und Elektrotechnik machen nur 3,3% keine Angaben bzw. haben kein Forschungsprojekt in den Jahren 1997 bis 1999 durchgeführt. Eine Forschung ohne Drittmiteleinsatz findet in den Ingenieurwissenschaften also praktisch nicht statt.

Abb. 12: Physik: Fachbereiche nach Drittmitteln und Publikationen  
(Blasenfläche = Drittmittel absolut)

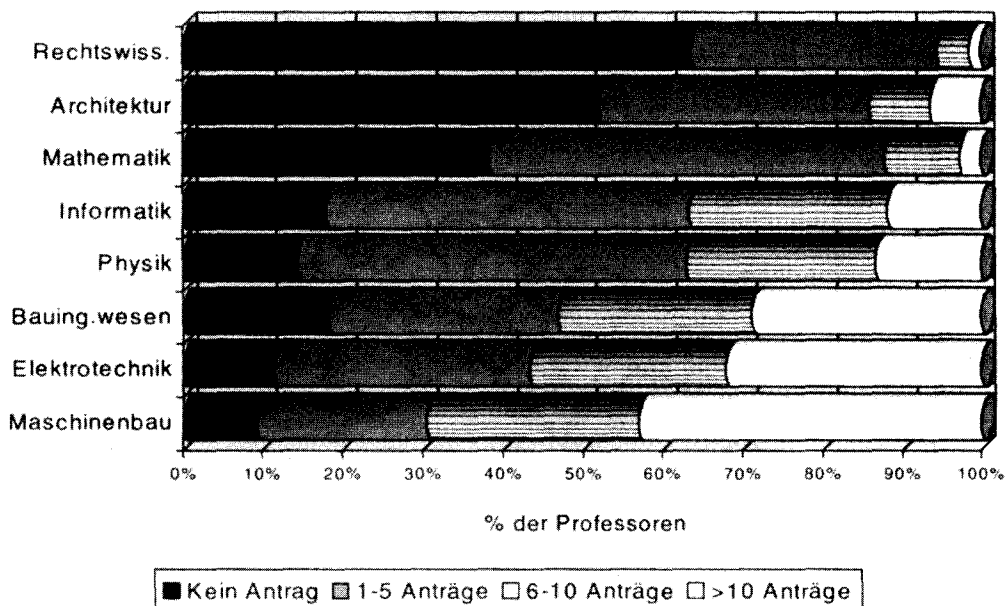


Voraussetzung für die Nutzung eines Drittmittelindikators ist, dass Drittmittelforschung in den untersuchten Fachgebieten üblich ist und nicht ausschließlich im Rahmen subdisziplinärer Spezialisierung (insbes. empirische Forschung) betrieben wird. Wie der Abbildung 13 zu entnehmen ist, sind die Rechtswissenschaften und die Architektur fast drittmittelabstinent. Aus diesem Grund macht es z.B. keinen Sinn, die Drittmittelinwerbungen in der Rechtswissenschaft als Indikator für Forschungsleistungen zu benutzen.

Auch die Nutzung der verschiedenen Finanzierungsquellen fällt disziplinspezifisch aus. Wie Abbildung 13 zeigt, haben die kaum Drittmittelforschung betreibenden Juristen keine Schwerpunkte bei der Wahl ihrer Finanziere. Physiker und Mathematiker hingegen stellen ihre Drittmittelanträge vornehmlich bei der DFG. Die Ingenieurwissenschaften schließlich akquirieren ihre Mittel überwiegend in der Privatwirtschaft (vgl. dazu den Beitrag von Güdler in diesem Band). Da diesen Mitteln sehr unterschiedliche Antragsbegutachtungen zugrunde liegen, stellt sich auch die Frage, ob allen Drittmittelzuflüssen eine ähnliche Indikatorqualität für Forschungsleistungen attestiert werden kann (vgl. Hornbostel 2001b).

Ähnlich wie für die Drittmittelinwerbungen gilt für die Publikationen, die man zunächst einmal als Aktivitätsindikator im Forschungsbereich interpretieren kann, dass disziplinspezifische Besonderheiten zu berücksichtigen sind: Wie die Abbildung 14 zeigt, ist in der Physik mit Artikeln in Fachzeitschriften und Kongressbeiträgen der ganz überwiegende Teil der Publikationen erfasst. Zudem ist die Reputation der verschiedenen journals nicht sonderlich umstritten, so dass z.B. mit dem Science Citation Index oder spezialisierteren Fachdatenbanken relativ problemlos „wichtige“ Beiträge identifiziert werden können.

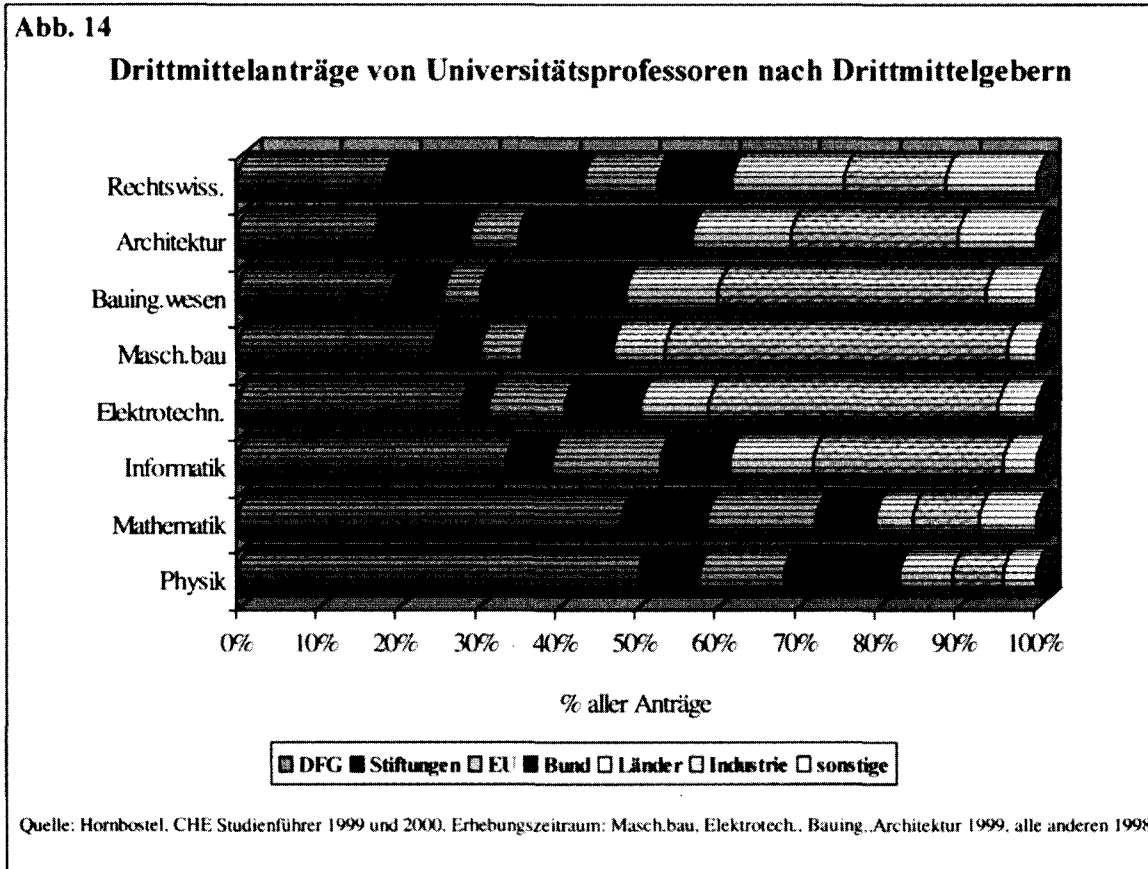
**Abb. 13** Drittmittelanträge von Universitätsprofessoren in den letzten 3 Jahren  
(Prozent der befragten Professoren)



Quelle: Hornbostel, CHE Studienführer 1999 und 2000  
Erhebungszeitraum: Masch.bau, Elektrotechnik, Architektur 1999, alle anderen 1998

In den Ingenieurwissenschaften spielt sich ein ganz erheblicher Teil der wissenschaftlichen Kommunikation in Kongressbänden ab. Ganz anders die Situation bei den Rechtswissenschaft-

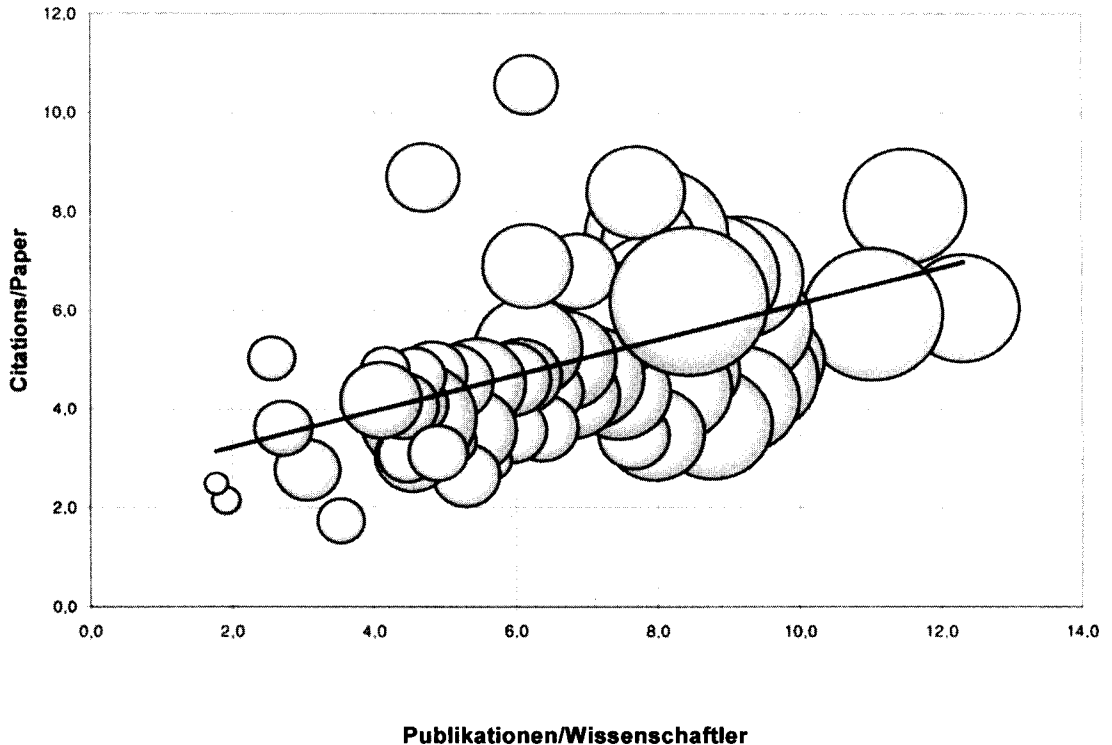
lern: Dort spielen Festschriften (Sammelbände) und Monographien und als Reflex darauf Rezensionen eine wichtige Rolle. Artikel in Fachzeitschriften streuen über die unterschiedlichsten journals (auch nicht-juristische). Die Unterscheidung zwischen originärer Forschung, Kommentierung und Dokumentation ist schwer zu ziehen. Über die Wertigkeit einzelner Zeitschriften herrscht wenig Einverständnis. Wollte man die juristischen Publikationen zu einem Indikator verarbeiten, wäre es nicht nur notwendig, die unterschiedlichen Publikationstypen auf irgendeine Weise zu verrechnen, sondern auch den diversen Zeitschriften Gewichte zuzuordnen. Im CHE Studienführer werden daher jeweils fachspezifische Publikationsindikatoren gebildet und auch nur dann publiziert, wenn diese Indikatoren aussagekräftig sind.



Publikationszählungen wie im Studienführer des CHE (Physik), die auf dem Science Citation Index beruhen, enthalten bereits eine Qualitätskomponente, denn nur weltweit stark zitierte Zeitschriften, werden in diese Datenbank aufgenommen. Allerdings erzeugen auch in vielzitierten Zeitschriften die einzelnen Beiträge eine sehr unterschiedliche Resonanz (Zitationen). Für eine Berichterstattung über Forschungsleistungen wird man auch diese Qualitätseinschätzung berücksichtigen müssen. Möglich ist das jedoch nur dort, wo auch ein hinreichender Anteil der Publikationen im Science Citation Index erfasst ist, das heißt überwiegend in den Naturwissenschaften. Abbildung 15 zeigt anhand der Daten des CHE Studienführers, dass zwischen Publikationsintensität und der durchschnittlichen Resonanz (erhaltene Zitate) der Artikel zwar eine positive Beziehung besteht, auch hier zeigen sich aber gerade im mittleren Bereich ganz erhebliche Unterschiede in der erreichten Resonanz in der Fachöffentlichkeit.

In Fachgebieten, für die Zitationsanalysen nicht möglich oder nicht sinnvoll sind, kann man derartige Qualitätsindikatoren entweder gar nicht nutzen oder nur auf Umwegen – über Gewichtungungsverfahren – Annäherungen versuchen.

**Abb. 15: Physik- Publikationen und Zitationen 1995-97 (Blasengröße = Publ. abs.)**

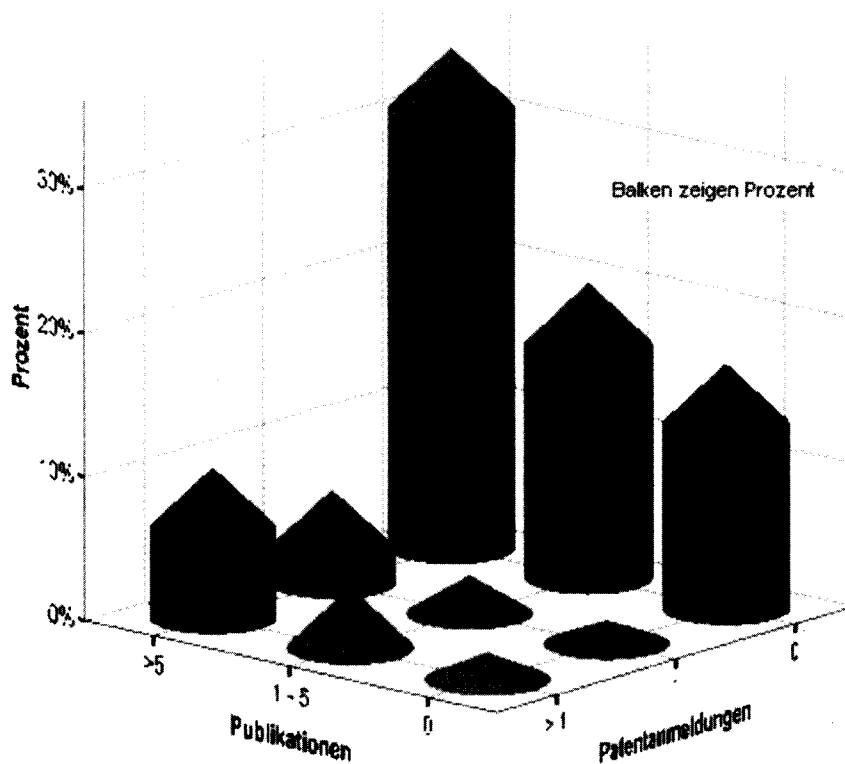


In stark anwendungsorientierten Disziplinen stellt sich ein zusätzliches Problem: Ein Teil der Forschungsarbeiten fließt nicht primär in das wissenschaftliche Kommunikationssystem, sondern in Anwendungskontexte, häufig im Rahmen von Kooperationen mit Wirtschaftsunternehmen. Einen Anhaltspunkt über das Ausmaß dieser innovativen, anwendungsorientierten Forschung lässt sich über die Zahl der Patentanmeldungen der Professoren gewinnen (vgl. dazu den Beitrag von Schmoch in diesem Band). Ein eigenständiger Indikator ergibt sich insbesondere deshalb, weil Publikationen patentschädlich sein können, insofern also ein Spannungsverhältnis besteht hinsichtlich der Sicherung von Prioritätsansprüchen. Anwendbar ist ein solcher Indikator aber nur in Disziplinen, in denen Patentanmeldungen regelmäßig und in einem erheblichen Umfang vorkommen. Das gilt etwa für den Maschinenbau, die Elektrotechnik, die Medizin und auch für die Chemie (vgl. Becher u.a. 1996).

Die Gewinnung entsprechender Daten ist ausgesprochen mühsam, da in den einschlägigen Datenbanken nicht ohne weiteres erkennbar ist, ob eine Patentanmeldung auf einen Hochschulprofessor zurückgeht oder nicht. Für die Patentanalyse in den Fächern Elektrotechnik und Maschinenbau des Studienführers des CHE wurde daher eine namentliche Überprüfung aller Hochschulprofessoren in der Patentdatenbank des Deutschen Patentamtes (PATDPA) durchgeführt. Ausgewählt wurden alle veröffentlichten deutschen Patent- und Gebrauchsmusteranmeldungen sowie alle veröffentlichten europäischen und PCT-Patentanmeldungen (Patent Cooperation Treaty der World Intellectual Property Organization) mit Priorität (bzw. prioritätsentsprechender Anmeldung) in den Jahren 1995 bis 1997. Dabei wurden mehrere Veröffentlichungen zu einer Priorität sowie eine Veröffentlichung zu mehreren Prioritäten jeweils nur einmal gezählt.

Abb. 16

**Patentanmeldungen und Publikationen in Fachzeitschriften 1996-98  
Universitätsprofessoren aus dem Fachgebiet Maschinenbau/Verfahrenstechnik**

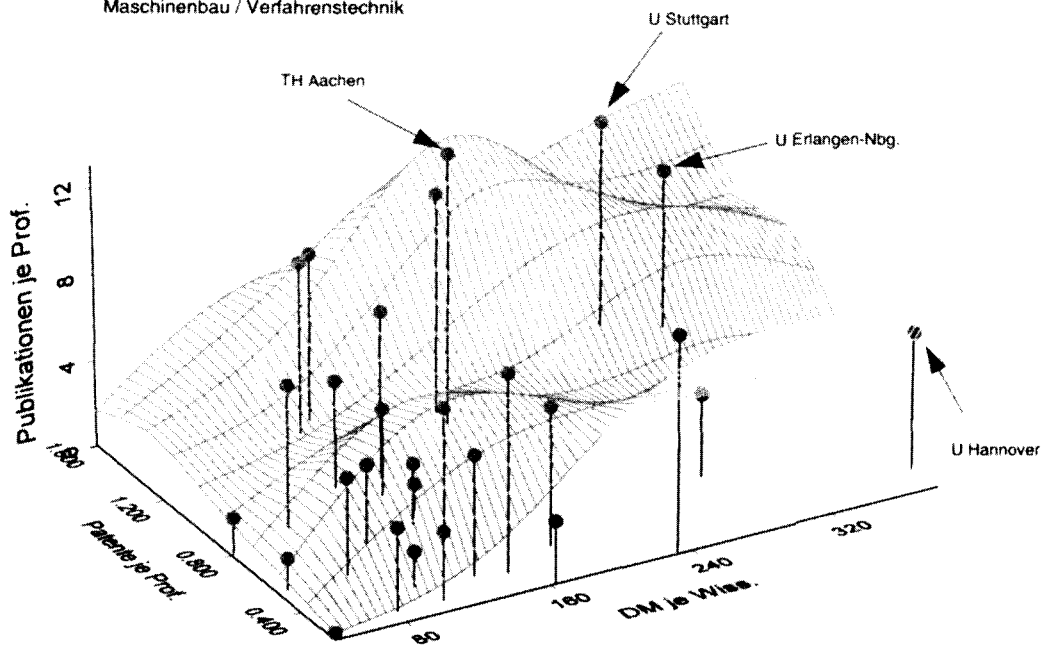


Im Ergebnis zeigt sich, dass nur ca. 28% der Professoren (Maschinenbau/Verfahrenstechnik an Universitäten) im Untersuchungszeitraum eine oder mehrere Patentanmeldungen eingereicht haben. Von diesen gehört allerdings nur eine Minderheit (ca. 13%) zu jenem Typus, der ausschließlich Patentanmeldungen vornimmt und nicht mit Publikationen auch im wissenschaftlichen Kommunikationssystem präsent ist. Wie Abbildung 16 zeigt, ist der größte Teil der patentaktiven Professoren auch mit Publikationen vertreten. Allerdings lassen sich für immerhin 17% der Maschinenbauprofessoren (nur Universitäten) weder Patentanmeldungen noch Publikationen im Untersuchungszeitraum nachweisen.

Angesichts dieser Fülle von Einzelinformationen stellt sich – ähnlich wie für die Lehrindikatoren – die Frage, wie man einerseits den fachlichen und innerfachlichen Differenzierungen mit aussagekräftigen Indikatoren gerecht werden kann und andererseits auch für den Nicht-Spezialisten einen orientierenden Überblick über die Forschungsleistungen bieten kann. Abbildung 17 vereinigt drei ausgewählte Indikatoren (Publikationen, Patente und Drittmittel). Erkennbar ist, dass nur sehr wenige Hochschulen auf ausschließlich einem Indikator hohe Werte erreichen, in der Regel harmonisieren die verschiedenen Messungen durchaus. D.h. die einzelnen Messungen unterschiedlicher Aspekte der Forschungsleistung gehen zwar nicht in einander auf, korrelieren aber deutlich untereinander. Diese Übereinstimmung lässt sich für einen weniger differenzierten, orientierenden Überblick nutzen.

Abb. 17

Drittmittel je Wissenschaftler, Patentanmeldungen je Prof., Publikationen je Prof.  
 Maschinenbau / Verfahrenstechnik



Zugleich erlaubt eine solche Zusammenfassung die Minimierung von Messfehlern einzelner Indikatoren, die kaum vermeidbar sind. Sie ist im übrigen auch kompatibel mit den Reputationsurteilen, die die Professoren über jeweils anderen Hochschulen abgeben (vgl. oben, Tab. 2). Der Unterschied besteht darin, dass der Hochschullehrertipp nur jeweils eine sehr kleine Gruppe von Hochschulen erfasst, während die Forschungsindikatoren zwar Leistungsunterschiede in der gleichen Richtung signalisieren, aber erstens mit weniger starkem Gefälle zwischen Hochschulen und zweitens auch mit markanten Abweichungen zwischen der Bekanntheit und dem Leistungspotential. Abb 18 zeigt diese Zusammenhänge für die im CHE Studienführer untersuchten ingenieurwissenschaftlichen Studiengänge. Besonders stark empfohlene Hochschulen (wie etwa TH Stuttgart oder TH Aachen) weisen auch hohe Werte für Publikationen und Patentanmeldungen pro Professor auf, andere Hochschulen (wie z.B. UGH Paderborn, Maschinenbau oder Universität Ulm, Elektrotechnik) weisen zwar ebenfalls hohe Publikationszahlen und Patentanmeldungen pro Professor auf, werden aber von relativ wenigen Professoren als Studienort empfohlen. Reputation, wie sie sich in den Studienortempfehlungen der Professoren ausdrückt, verändert sich nur langsam.

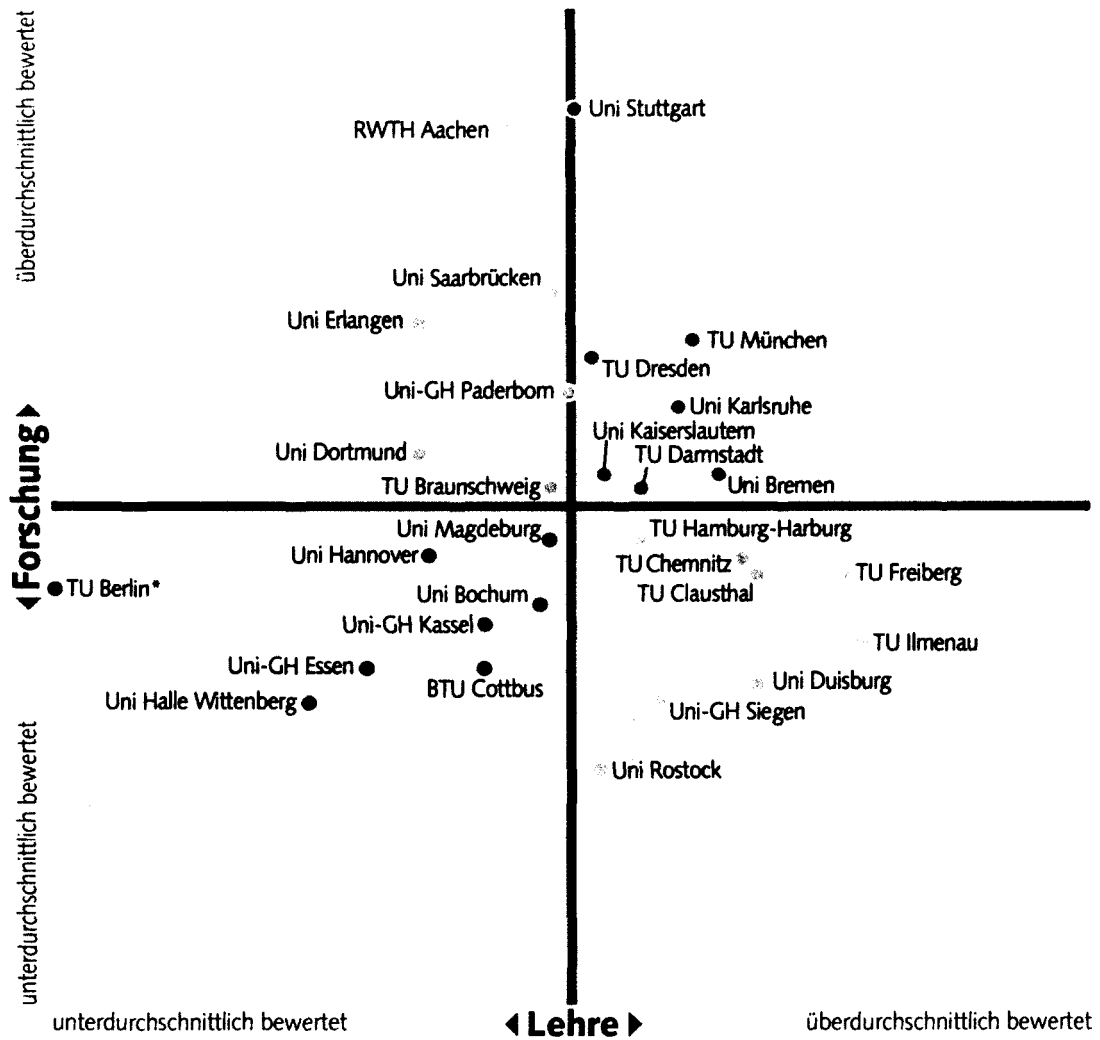
### 10. Einwand: Rankings fehlt die Zielgruppenorientierung

Das Grundmodell für diese Forderung ist im Beitrag von Bayer in diesem Band skizziert. Es geht von einer Marketingstrategie aus und bemüht dazu Analogien aus dem Wirtschaftsbereich. Abiturienten oder Studierenden sind wie die Käufer von Autoreifen mit bestimmten Präferenzen ausgestattet (von sportlich bis sparsam) und sehen sich einem differenzierten Marktangebot gegenüber. Die Bewertung der Produkte hängt vollständig an den Präferenzen der Käufer. D.h. der gleiche Autoreifen wird von den einen als sehr gut von den anderen als sehr schlecht be-

wertet. Aus diesem Modell wird nun einerseits hergeleitet, dass Rankings nicht aussagekräftig seien, weil auch dann, wenn der Prozentsatz von sparsamen Käufern und sportlichen gleich groß ist, keine entscheidungsunterstützende Information entstünde, andererseits wird daraus die Forderung abgeleitet, adressatenspezifische Kriterienkataloge zu entwickeln. Das erste Argument ist schlicht und logisch falsch: Würden die sportlichen Käufer die Billigreifen immer schlecht bewerten, dann käme bei einer Mittelwertbildung die Urteilsdifferenzierung der sparsamen Käufer zum tragen, würden die sparsamen alle hochwertigen und teuren Reifen schlecht bewerten kämen die Urteilsdifferenzierung der Sportwagenfahrer zum tragen. Bei gegebenen Präferenzen wäre genau die entscheidungsrelevante Information entstanden, die gesucht wurde. Schlicht ist das Argument, weil eine Studienentscheidung nun einmal unter etwas komplexeren Bedingungen verläuft als ein Reifenwechsel. Die empirisch wichtigste Differenzierung auf der Käuferseite ist die Unterscheidung von akademischen und praktischen Präferenzen. Diese Differenz ist aber schon in den Institutionen abgebildet (Fachhochschule, Universität). Um im Bild zu bleiben: Nicht die Urteile aller Reifenkäufer werden gemittelt, sondern getrennt, die der Kunden spezialisierter Reifenhändler. Abgesehen davon, dass sich ein Kunde gelegentlich verirrt, sind damit die wichtigsten Präferenzunterschiede bereits berücksichtigt, denn niemand hat bisher Universitäts- und Fachhochschulranglisten durcheinandergewürfelt. Die verbleibenden Bewertungsunterschiede bei den Studierenden folgen konsequenterweise auch nicht ausgeprägten Urteilsprofilen, sondern stellen überwiegend Niveaushiftungen dar, so dass in den meisten Fächern eine Aufbereitung der Daten für einzelne Gruppen (vgl. die Clusteranalyse) die Zahl der Indikatoren explosionsartig wachsen lässt, den Informationsgehalt (Rangliste) jedoch kaum verändert. Tod durch Informationsentropie wäre die Diagnose für ein solches Ranking angesichts des Bedarfs an einfachen Orientierungsgrößen, die Abiturienten in Gruppendiskussionen über den CHE-Studienführer einfordern. Auch aus einem weiteren Grund ist die von Bayer geforderte Zielgruppenorientierung wenig erstrebenswert. Es geht beim Ranking nämlich nicht um Marketing – auch wenn viele Hochschulen die Ergebnisse für diese Zwecke nutzen –, sondern um die Bereitstellung entscheidungsrelevanter Informationen. Was entscheidungsrelevant ist, sollte man aber mündigen Konsumenten durchaus selbst überlassen. Die Degradierung zur „Zielgruppe“, die dann einen Indikatorencocktail bekommt, den die Evaluateure für angemessen halten, ist aus einer Marketingperspektive erstrebenswert, aber nicht aus der Perspektive von Information und Orientierung suchenden Interessenten. Der CHE Studienführer in der Online-Version überlässt daher den Informationssuchenden die Kombination von Kriterien für eine Auswahl, darunter auch Kriterien, die nichts mit der Qualität der Ausbildung zu tun haben, sondern mit sozialen oder finanziellen Belangen.

Die Forderung nach immer weiteren Differenzierung beruht auf einem Missverständnis der Funktion von Rankings. Diese ersetzen nämlich keineswegs die Evaluation an den Hochschulen oder die Studienberatung. In hochschulinternen Evaluationen kann eine detaillierte Analyse der Zusammenhänge zwischen Erwartungen und Voraussetzungen der Studierenden und deren Beurteilungen zu steuerungsrelevantem Wissen führen (vgl. Krempkow 1999). Ortsvergleichende Evaluationen hingegen müssen einen Spagat bewältigen zwischen einer möglichst hohen Komprimierung von Informationen und einer Differenzierung und Detaillierung für Informationssuchende, die Vertiefung bereits vorhandenen Wissens suchen.

# Faktorenanalyse Maschinenbau



\* Der Faktorwert für den Faktor Lehre liegt bei der TU Berlin außerhalb des Darstellungsbereichs

**Abb. 18**

Rotierte Komponentenmatrix \*

	Faktor	
	Forschung	Lehre
Publikationen je Professor **	.830	
Studienortempfehlung Professor	.739	
Drittmittel je Professor	.591	
Patentanmeldungen je Professor	.589	
Studierendenurteil Lehre		.825
Professorenurteil Lehre	-.112	.810

\* Hauptkomponentenanalyse, Varimax Rotation mit Kaiser-Normalisierung \*\* mit Autorenzahl gewichtet

Fehlende Werte durch Mittelwerte ersetzt, Erklärter Varianzanteil: Faktor Forschung 33%, Faktor Lehre 22%

Ein Beispiel für eine Informationsverdichtung, die auch auf Ranggruppenbildung verzichtet, sind die im CHE Studienführer 2000 benutzten Faktorenanalysen über die Hauptdimensionen Forschung und Lehre (vgl. Abb.18). Für diese Faktorenanalysen werden sowohl subjektive Urteile von Professoren und Studierenden als auch „harte Fakten“ (Forschungsindikatoren) verrechnet. Übrig bleiben die „Konsenszonen“ verschiedener Beurteilungen. Damit ist ein Ausgangspunkt gegeben für eine Informationserschließung, die zumindest in einem elektronischem Medium auch sehr differenzierte und detaillierte Datenpräsentationen erlaubt. Anstelle von vordefinierten Zielgruppen entsteht auf diese Weise ein orientierendes Informationssystem, das Fakten ebenso wie die subjektiven Urteile unterschiedlicher Bewertergruppen (und wo geboten auch von Subpopulationen) bereitstellt.

**Literatur:**

- Becher, G., Gering, T., Lang, O. / Schmoch, U. Patenwesen an Hochschulen. Hrsg. Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie. Bonn 1996.
- Daniel, H.-D. / Hornbostel, S. „Das SPIEGEL-Ranking. Mediensensation oder ein Beitrag zur hochschulvergleichenden Lehrevaluation?“ Universität und Lehre. Hrsg. P. Mohler. Münster 1994: 29-44.
- Güdler, J., Sack, D. „Berufsfeld Wissenschaft. Zum Einfluss institutioneller Reputation auf die Platzierung von Nachwuchs-Soziologen“. Soziologielehre in Deutschland. Hrsg. H.M. Artus und M. Herfurth. Opladen 1996: 143-178.
- Hornbostel, S. / Daniel, H.-D. „Studienbedingungen in der Soziologie“. Soziologielehre in Deutschland. Hrsg. H.M. Artus und M. Herfurth. Opladen 1996: 11-58.
- Hornbostel, S. Wissenschaftsindikatoren. Bewertungen in der Wissenschaft. Opladen 1997.
- Hornbostel, S. „Der Uni-Test Europa des SPIEGEL: Infotainment oder Entscheidungshilfe?“ Uni-Test Europa. Wo sich das Studieren im Ausland lohnt. Hrsg. M. Doerry und J. Mohr. Hamburg 1998: 149 - 162.
- Hornbostel, S. „Das SPIEGEL-Ranking deutscher Hochschulen und die Folgen: Interaktionsprozesse zwischen Öffentlichkeit und Wissenschaft“. Die Eigenwilligkeit sozialer Prozesse Hrsg. J. Gerhards und R. Hitzler. Opladen 1999: 174-205.
- Hornbostel, S. „Evaluation und Ranking - Führen sie zu mehr Transparenz und Vergleichbarkeit?“ Beiträge zur Hochschulpolitik. Hrsg: Hochschulrektorenkonferenz. 4 1999: 81-96.
- Hornbostel, S. „Welche Indikatoren zu welchem Zweck: Input, Throughput, Output.“ Qualitätsförderung durch Evaluation? Ziele, Aufgaben und Verfahren im Wandel Hrsg. M. Röbbcke und D. Simon. Wissenschaftszentrum Berlin 1999: 99-103.
- Hornbostel, S. „Der Studienführer des CHE – ein multidimensionales Ranking“. Hochschulranking. Zur Qualitätsbewertung von Studium und Lehre. Hrsg. U. Engel. Frankfurt/M 2001a: 83-120.
- Hornbostel, S. „Third Party Funding of German Universities. An Indicator of Research Activity?“ Scientometrics Vol. 50 2001b: 523-537.
- Koch, J. „Wenn „mehr“ nicht gleichbedeutend mit „besser“ ist: Ausschöpfungsquoten und Stichprobenverzerrungen in allgemeinen Bevölkerungsumfragen“. ZUMA Nachrichten 42 1998: 66-90.
- Krempkow, R. Ist „gute Lehre“ meßbar? Untersuchungen zur Validität, Zuverlässigkeit und Vergleichbarkeit studentischer Lehrbewertungen. Marburg 1999.
- Kromrey, H. „Qualität und Evaluation im system Hochschule“. Evaluationsforschung (Hrsg. R. Stockmann). Opladen 2000.
- Lewin, K., Heublein, U., Schreiber, J., Sommer, D. „Vorbereitung auf das Studium und Informationsstand deutscher Studienanfänger bei Studienbeginn“. HIS Kurzinformation A8/1997.
- Meinefeld, W. „Hochschulranking. Eine unsichere Basis für Entscheidungen“. Forschung & Lehre 1 2000: 23-25.

Müller-Böling, D. / Hornbostel, S. „Fehlinterpretationen und Vorurteile. Vom Umgang mit Hochschulrankings und deren Nutzen“. *Forschung & Lehre* 2 2000: 81-38.

Ott, R. „Darstellung und Beurteilung von Hochschul-Rankings in Deutschland“. *Beiträge zur Hochschulforschung* 4 1999: 309-323.

Scheuch, E. K. „Wie gut sind unsere Hochschulen?“ *Soziologie* 2 1990: 73-90.

Schimank, U. *Hochschulforschung im Schatten der Lehre*. Frankfurt/M. 1995.

Süllwold, F. „Ranking ist oft ein Synonym für Unsinn“. *Forschung & Lehre* 11 1997: 578-579.

Weingart, P. „Forschungsindikatoren: Instrumente politischer Legitimierung oder organisatorischen Lernens“. *Qualitätssicherung in Hochschulen*. Hrsg. Detlef Müller-Böling. Gütersloh 1995: 73-84.