Leuchttürme oder Scheinriesen

Wie zuverlässig sind die Hochschulrankings der Massenmedien?[1]

Uta Liebeskind
Wolfgang Ludwig-Mayerhofer
Siegen

Mit dem SPIEGEL, dem FOCUS und der ZEIT legen große deutsche Wochenzeitschriften seit einigen Jahren regelmäßig vergleichende Rankings von Hochschulen vor mit dem Ziel und dem Anspruch, Orientierung in der deutschen Hochschullandschaft zu verschaffen.[2] Die Rankings sind Ausdruck der in den 90er Jahren entdeckten Unterschiedlichkeit deutscher Hochschulen und insofern konsequente Begleiterscheinung des Versuches, nun auch die politisch gewollte Profilierung von Hochschulen zu etablieren. Die Bemühungen um die Herausarbeitung der Differenzen zwischen Hochschulen sind nicht nur auf die Medien beschränkt: der Wissenschaftsrat etwa, wichtiges Beratungsgremium für bildungspolitische Entscheidungen im Hochschulbereich, hat bereits vor zwei Jahren Empfehlungen für die Erstellung von Rankings im Wissenschaftsbereich abgegeben – zunächst allerdings auf Forschung beschränkt (Wissenschaftsrat 2004).

[1] Wir danken Lena Ellenberger und Christian Hoffmann für die Eingabe der umfangreichen Ranking-Daten, die sie mit Geduld und viel Sorgfalt vorgenommen haben.

[2] „Das Ranking des Centrums für Hochschulentwicklung (CHE) hilft Schülern und Studienwilligen, die richtige Hochschule zu finden." (http://www.das-ranking.de/che7/CHE; Zugriff am 23.05.2006); „Deutschlands beste Uni" benennt das FOCUS-Ranking (http://bildung.focus.msn.de/bildung/bildung/unilisten; Zugriff am 23.05.2006).

Die mittlerweile scheinbar selbstverständliche Notwendigkeit von Rankings macht eine kritische Auseinandersetzung mit ihnen umso dringlicher. Ein kritischer Blick muss mehrere Perspektiven haben. Zunächst ist zu fragen, ob Rankings in Deutschland grundsätzlich mehr Licht ins Hochschulsystem bringen können. Hier ist Skepsis angebracht. Rankings sind Instrumente zur Orientierung in marktförmig organisierten Systemen, sie dürften somit ein Fremdkörper im deutschen Hochschulsystem sein (Pechar 1997). Ein Vergleich mit den USA mit ihrem schier unüberschaubaren Angebot sehr unterschiedlicher Studienmöglichkeiten zeigt die viel geringere Marktförmigkeit des deutschen Hochschulwesens auf. Die deutschen Universitäten werden aus staatlichen Mitteln finanziert; die angebotenen Bildungsgänge waren bis vor kurzem zwischen einzelnen Hochschulen nur sehr wenig differenziert. Rankings haben hierzulande somit etwas Künstliches: Sie messen qualitative „Unterschiede", die mit dem Blick auf andere Hochschulsysteme – jedenfalls bislang – kaum der Rede wert sind.

Zweifel an der Orientierungsfunktion bewertender Rankings für zukünftige Studierende – die von den Medien als eigentliche Adressaten der Rankings angesprochen werden – kommen auch auf, wenn man den Prozess-Charakter des *Lernens* ernst nimmt (Mächtle und Witthaus 2002). Zunächst ist bei einigen Indikatoren, die als Indikatoren für die Qualität der Lehre angeboten werden, durchaus fraglich, ob sie hierüber überhaupt etwas auszusagen geeignet sind; hier ist etwa an die Reputation der Lehre bei Professoren oder bei Unternehmern zu denken, aber auch an die (vom FOCUS angebotenen) Studiendauern. Aber auch dort, wo ein Bezug zur Lehrqualität bestehen könnte – wie etwa im Falle subjektiver Studierendenurteile oder der (freilich auf dem Papier konstruierten) Betreuungsquoten –, gehen die Rankings darüber hinweg, dass angesichts heterogener Interessen und ganz unterschiedlicher Modi studentischer Aneignung von und Auseinandersetzung mit Wissen durchschnittliche Angaben kein Urteil darüber erlauben, ob eine individuelle Studierende an einer bestimmten Universität Bedingungen vorfindet, die ihr entgegenkommen oder nicht.

Will man sich diesen Argumenten nicht oder nicht vollständig anschließen und den Rankings im deutschen Hochschulsystem nicht von vorneherein jeglichen Orientierungsnutzen absprechen, bleibt dennoch die Frage: Können die von uns betrachteten Rankings der Wochenzeitschriften FOCUS, SPIEGEL und ZEIT ihren Leserinnen verlässlich

Orientierung bei der Wahl ihrer Universität bieten, und zwar unabhängig davon, in welcher der Zeitschriften sie gerade blättern? Diese Frage ist ganz zentral; schließlich kann nicht davon ausgegangen werden, dass alle Nutzer die Qualität der konsultierten Rankings einschätzen können. Die allermeisten Nutzerinnen müssen sich vielmehr auf Seriositätsversprechungen wie „Sonderauswertungen des statistischen Bundesamtes" (FOCUS, Heft 36/2005: 103) oder „knapp 50000 Fragebögen analysiert" (SPIEGEL, Heft 48/2004: 181) verlassen. Wenn unterschiedliche Rankings zu ähnlichen Ergebnissen hinsichtlich der Bewertung der Universitäten kämen, dann wäre dies (als Reliabilität der Messungen) zwar kein Nachweis der Gültigkeit, könnte aber doch zumindest plausibel erscheinen lassen, dass den Rankings wenig Beliebigkeit anhaftet und sie somit als Orientierungshilfe bei der Wahl des Studienortes tauglich sein könnten. Die entscheidende Frage im Zusammenhang mit der Vielfältigkeit des Ranking-Angebotes lautet also: Stimmen die Ergebnisse der betrachteten Rankings im Grundsatz überein?

Bereits vor einem Jahr haben wir anhand unseres eigenen Faches, der Soziologie, festgestellt, dass von Übereinstimmung der drei betrachteten Rankings nicht die Rede sein kann (Liebeskind und Ludwig-Mayerhofer 2005); eine Überprüfung anhand einiger weniger weiterer Fächer ergab, dass es in den meisten anderen Fächern etwas besser aussieht, wenngleich kein Einzelindikator als durchgängig verlässlich erschien.[3] Ziel dieser Arbeit ist es nun, sämtliche Fächer, zu denen (mindestens) zwei Rankings vorliegen, hinsichtlich ihrer Bewertung durch die drei prominentesten Rankings zu vergleichen, um ein Gesamturteil über deren Übereinstimmung zu ermöglichen.

Ranking-Daten und Übereinstimmungsmaße

Wir vergleichen die jeweils aktuellsten Rankings, die der FOCUS, die ZEIT (in Kooperation mit dem Centrum für Hochschulentwicklung CHE)[4] und der SPIEGEL veröffentlicht haben. Die jüngsten veröffent-

[3] Auf die unterschiedlichen Ergebnisse von Rankings weist auch Meinefeld (2000: 27) hin, allerdings nur exemplarisch anhand weniger Fächer und ohne Verwendung von Kennztahlen für das Ausmaß der (Nicht-)Übereinstimmung.

[4] Wir sprechen im Folgenden vom CHE- und nicht vom ZEIT-Ranking, da zwar die ZEIT das zentrale Medium der Veröffentlichung des Rankings ist, Erhebung und Auswertung der

lichten Daten von FOCUS und vom CHE stammen aus 2005,[5] das letzte SPIEGEL-Ranking wurde bereits 2004 veröffentlicht.[6] Ein Vergleich aller drei Rankings kann jedoch nur hinsichtlich der globalen Einschätzung der Fächer angestellt werden, da sich das SPIEGEL-Ranking in Herangehensweise und Methodik sehr stark von den beiden anderen unterscheidet. Während FOCUS und CHE vergleichende Daten zu Fachbereichen gesammelt und ausgewertet haben,[7] erhebt der SPIEGEL in einer Online-Befragung Selbstauskünfte von Studierenden und aggregiert diese auf Fächerebene zu Ranglisten.[8] Einzig das CHE bewertet auch Fachhochschulen, die mangels Vergleichs in den anderen Rankings hier nicht berücksichtigt werden. Der FOCUS bezog reine Lehramtsstudiengänge nicht in die Erhebungen ein, das CHE differenzierte bei der Ergebnispräsentation zwischen Lehramtsstudiengängen und reinen Fachstudiengängen. Wir beziehen unsere Auswertungen daher nur auf letztere.[9]

Die Rankings basieren in Folge unterschiedlicher Datenerhebungsstrategien auf unterschiedlichen Fallzahlen (hier: Universitäten). Der SPIEGEL, der nur Universitäten berücksichtigt, an denen mindestens acht der 15 von ihm gerankten Fächer vertreten sind, weist durchgängig die geringsten Fallzahlen auf; vier der von den anderen Zeitschriften gerankten Fächer wurden vom SPIEGEL gar nicht berücksichtigt (Angli-

---

Daten jedoch im Gegensatz zu FOCUS und SPIEGEL gänzlich bei einer externen Institution, dem CHE liegen.

[5] Der FOCUS veröffentlichte sein Ranking in einer sechsteiligen Serie, beginnend mit Heft 36/2005. Das CHE veröffentlichte seine Ergebnisse in der Wochenzeitung DIE ZEIT beginnend mit Ausgabe Nr. 21 (19.05.2005), kompakt zusammengefasst in einem „Studienführer DIE ZEIT Ausgabe 2005/06“ und schließlich auf den Web-Seiten der ZEIT im Internet. Neueste CHE-Daten erschienen während des Verfassens dieses Artikels Anfang Mai 2006 (http://www.das-ranking.de/che7/CHE; Zugriff am 31.05.2006). Die neuen Daten konnten zu diesem Zeitpunkt mit Ausnahme einzelner Datenreihen zur Ergänzung fehlender Daten (siehe Tabelle 3) nicht mehr berücksichtigt werden.

[6] Der SPIEGEL veröffentlichte sein Ranking in Heft 48/2004.

[7] Zur Methodik des CHE siehe: http://www.che.de/downloads/CHE_HochschulRanking_Methoden2005.pdf (Zugriff am 02.06.2006); zur Methodik des FOCUS siehe: http://focus.msn.de/bildung/bildung/hochschulen (Zugriff am 02.06.2006)

[8] Zur Methodik des SPIEGELs siehe: http://www.studentenspiegel1.de/methodik.pdf (Zugriff am 02.06.2006)

[9] Das CHE differenziert mitunter innerhalb eines Faches zwischen verschiedenen Studiengängen (z.B. „Nordamerikastudien“ und „Englische Philologie“ im Fach Anglistik); diese behandeln wir als einzelne Fälle und vergleichen sie *jeweils* mit den auf das gesamte Fach bezogenen FOCUS-Daten.

stik, Bauingenieurwesen, Geschichte, Pädagogik). CHE und FOCUS betrachten im Durchschnitt mit jeweils knapp 50 Fällen fast doppelt so viele Universitäten wie der SPIEGEL.

Wir betrachten drei Gruppen von Indikatoren: Zunächst werden Globalurteile verglichen, die zum Ziel haben, so etwas wie die „Gesamtqualität" des Faches an den einzelnen Universitäten abzubilden. Das CHE verzichtete bewusst auf eine derartige Globaleinschätzung. Nutzer des CHE-Rankings könnten allerdings das in sämtlichen Übersichten an erster Stelle aufgeführte Gesamturteil der Studierenden als ein solches Globalurteil lesen; wir vergleichen dieser Lesart entsprechend diese Gesamturteile mit den Globalbewertungen der anderen beiden Rankings. Weiterhin betrachten wir Indikatoren zur Lehrqualität der Fächer, nämlich Daten zur Betreuungsrelation, zur Betreuungsqualität und zur Reputation der Lehre. Schließlich gehen wir auf die Übereinstimmung der Bewertung der Forschungsleistung im Fach ein. Hier liegen neben Angaben zur Forschungsreputation vor allem „harte" Fakten vor: Drittmittel, Promotionsquoten sowie Publikationen und/oder Zitationen.

Die Globalurteile und Reputationsmerkmale liegen als Rangdaten (also ordinalskaliert) vor: Alle drei Rankings ordnen die einzelnen Universitäten bzw. Fachbereiche (freilich mit jeweils eigenen Verfahren der Gruppenbildung) einer Spitzen-, Mittel- und Schlussgruppe zu. Alle anderen Angaben sind größtenteils metrisch skaliert, machen also (jedenfalls dem Anschein nach) Aussagen nicht nur darüber, ob, sondern auch um welchen Betrag eine Universität im betrachteten Merkmal besser oder schlechter ist als eine zweite.[10] Die Übereinstimmung der Rankings bestimmen wir grundsätzlich anhand von Korrelationsmaßen, die den Zusammenhang zwischen je zwei Merkmalen aus unterschiedlichen Rankings quantifizieren; für ordinalskalierte Merkmale wählen wir aus den möglichen Alternativen Kendalls Tau_b, für metrisch skalierte Merkmale den Pearsonschen Korrelationskoeffizienten.[11] Sowohl Kendalls Tau_b als

[10] In einigen Fällen hat der FOCUS die Daten zu Zitationen nur in gruppierter Form (Spitzen-, Mittel-, Schlussgruppe) veröffentlicht. In diesen Fällen werden zum Vergleich beim CHE ebenfalls die gruppierten Daten herangezogen.

[11] Die häufig vorgeschlagenen und verwendeten Maße Kappa (für ordinalskalierte Daten) bzw. ICC (für metrische Daten) sind für unsere Zwecke weniger brauchbar: Die Eignung von Kappa zur Messung der Übereinstimmung der Einstufung von Objekten ist in der wissenschaftlichen Literatur äußerst umstritten (Uebersax 1987; Guggenmoos-Holzmann 1993); bei ordinalskalierten Daten hängen die Ergebnisse obendrein von der letztlich willkürlichen Wahl von Gewichten ab (Brenner/Kliebsch 1996). Die Intraklassenkorrelation

auch der Pearsonsche Korrelationskoeffizient können Werte im Bereich von –1 bis 1 annehmen; Werte nahe 1 (bzw. –1) kennzeichnen perfekte positive (bzw. negative) Zusammenhänge, Werte nahe 0 zeigen an, dass kein Zusammenhang besteht.

Bei der Bestimmung des Pearsonschen Korrelationskoeffizienten wurde jeweils anhand von Streudiagrammen überprüft, ob einzelne Fälle die Korrelation deutlich positiv oder negativ beeinflussten; war dies der Fall, so berichten wir auch das Korrelationsmaß, das nach Elimination des betreffenden Falls bzw. (vereinzelt) mehrerer Fälle ergab, sofern der Unterschied größer als |0,1| war.

Welche Maßstäbe sind hinsichtlich der errechneten Zusammenhänge anzulegen? Die Rankings entsprechen der psychologischen Falldiagnostik: Es werden Urteile über einzelne Fälle – hier: Universitäten – getroffen. Hier sollte die Reliabilität nach Bortz/Döring (2002: 199) mindestens 0,8 betragen; liegt sie zwischen 0,8 und 0,9 kann sie nach diesen Autoren als „mittelmäßig", erst ab einem Wert von 0,9 als „hoch" gelten. Lienert/Raatz (1994: 269) bezeichnen auch Werte von 0,7 als „eben noch ausreichend"; wir schließen uns dieser Auffassung an.

Globalurteil

Die Betrachtung der Globalurteile zeigt allenfalls mäßige Übereinstimmung der Rankings (Tabelle 1). Zwar ist zu sehen, dass die Koeffizienten überwiegend positive Werte annehmen, was nicht zu erwarten wäre, wenn es sich um reine Zufallsprodukte handelte.[12] Ihr durchweg kleiner Betrag zeigt jedoch eindrücklich, dass die Zuordnung zu Spitzen-, Mittel- und Schlussgruppe eben nicht übereinstimmend erfolgt. Vielmehr sind in den entsprechenden Kreuztabellen die Zellen jenseits der Diagonalen – letztere weist die kongruenten Urteile aus – teils recht stark besetzt; das Gros der Werte liegt zwischen 0 und nur 0,50, sieht man von den vier Fächern ab, die diesen Wert mehr (Maschinenbau, SPIEGEL

(ICC) ist heranzuziehen, wenn Unterschiede in Mittelwerten und/oder Metriken zwischen unterschiedlichen Einstufungen von Bedeutung sind (Wirtz/Caspar 2002: 157ff.); hier sind solche Unterschiede jedoch gerade irrelevant, denn es zählt (auch und gerade aus der Sicht der möglichen Nutzer von Rankings) nur die (relative) Platzierung, nicht der absolute Messwert.

[12] Dann müssten sich sowohl zufällige negative als auch zufällige positive Zusammenhänge zeigen, die im Mittel den Wert 0 aufweisen.

versus CHE) oder weniger (Bauingenieurwesen, FOCUS versus CHE und Elektrotechnik und Jura, SPIEGEL versus FOCUS) deutlich überschreiten.

Tabelle 1: Globalurteile im Vergleich, CHE, FOCUS und SPIEGEL (Kendalls Tau$_b$)

	SPIEGEL vs. FOCUS	SPIEGEL vs. CHE[(a)]	FOCUS vs. CHE[(a)]
Anglistik	–	–	.24
Bauingenieurwesen	–	–	.60
Biologie	.17	.03	.23
BWL	.47	.11	.16
Chemie	.28	.06	–.08
Elektrotechnik	.54	.15	.28
Germanistik	.46	.31	.17
Geschichte	–	–	.22
Informatik	.06	.06	.11
Jura	.53	.25	.07
Maschinenbau	.41	.86	.36
Mathematik	.35	.14	.01
Medizin	.50	.01	–.24
Pädagogik	–	–	.08
Physik	.35	.23	–.03
Politikwissenschaft	.44	.12	.19
Psychologie	.01	.26	.23
Soziologie	.24	.00	.28
VWL	.18	.25	.37

– In (mindestens) einem Ranking keine Daten vorhanden

(a) Gesamturteil der Studierenden

Möglicherweise ist für diejenigen, die Rankings zur Orientierung nutzen wollen, nur die übereinstimmende Identifizierung von „Spitzenuniversitäten“ von Interesse. Die Rankings wären dann aus der Nutzerperspektive verlässlich, wenn sich die Abweichungen nur in der Zuordnung zu Mittel- und Schlussgruppe ergeben würde. Doch auch hier ist Vorsicht geboten. Selbst wenn, wie z.B. im Fall der Rankings von CHE und

FOCUS für das Fach Bauingenieurwesen ein Wert von knapp 0,6 ein gewisses Maß an Übereinstimmung der Ergebnisse nahe legt, ist nicht ausgeschlossen, dass einzelne Universitäten von einem Ranking der Spitzen-, vom anderen jedoch der Schlussgruppe zugeordnet werden. Im Beispiel trifft das nur auf eine Universität zu, in einigen Fächern werden jedoch bis zu 11 (!) Universitäten in jeweils entgegen gesetzte Extremgruppen eingeordnet.

Tabelle 2: Indikatoren zur Lehre im Vergleich, CHE versus FOCUS

	Betreuung[(a)] Pearsons r[(b)]		Reputation Lehre Kendalls Tau_b
Anglistik	.29		.59
Bauingenieurwesen	.56		.74
Biologie	.20		.60
BWL	–		.47
Chemie	.09		.49
Elektrotechnik	–.03	(.31)	.66
Germanistik	.27		.62
Geschichte	.31	(.41)	.60
Informatik	–.03	(.12)	.58
Jura	.52	(.64)	.54
Maschinenbau	.07		.60
Mathematik	.07		.48
Medizin	–.14	(.59)	.57
Betreuungsrelation FOCUS/CHE	–.01	(.44)	–
Pädagogik	–		.60
Physik	.52		.57
Politikwissenschaft	.44		.50
Psychologie	.46		.34
Soziologie	.35		.64
VWL	–		.62

– In (mindestens) einem Ranking keine Daten vorhanden

(a) Betreuungsrelation (FOCUS) versus Studierendenurteil zur Betreuungssituation (CHE)

(b) Koeffizienten nach Elimination von Ausreißern in Klammern

Es fällt auf, dass die Übereinstimmung der Globalurteile zwischen FOCUS und SPIEGEL durchschnittlich am größten ist, obwohl das SPIEGEL-Ranking in seiner Vorgehensweise deutlich von den beiden anderen Rankings abweicht. Über die Gründe für die – dennoch auch nur vergleichsweise – hohe Übereinstimmung kann nur spekuliert werden. Zum einen treten an dieser Stelle möglicherweise Effekte einer self-fulfilling prophecy auf: Universitäten, denen z.B. von Rankings ein herausgehobener Status zugeschrieben wird, ziehen möglicherweise stärker als andere Universitäten Studierende an, die den Studienort eher nach fachbezogenen denn nach lebenspraktischen Kriterien auswählen. Unterstellt man, dass diese Studierenden dann eine stärkere Fachbindung aufweisen und dass dies auch mit besseren Leistungen verbunden ist, könnte sich tatsächlich der Effekt einstellen, dass sich im Fach engagierte Studierende an den in der Spitzengruppe befindlichen Universitäten ballen. Verstärkend könnte ein zweiter Effekt, ein Selektionseffekt wirken. Denkbar ist, dass eben jene „fachbewussten" Studierenden, die sich an einer Spitzenuniversität wähnen, ein stärkeres Interesse an der Teilnahme am SPIEGEL-Ranking haben, das Selbstpositionierung und Vergleich mit anderen Studierenden verspricht. Die Wahrnehmung von Konkurrenz und die Notwendigkeit, Exzellenz zu demonstrieren, könnten unter dieser Studierendengruppe stärker ausgeprägt sein.

Einzelindikatoren – Bewertung der Lehrsituation

Zur Beurteilung der Lehrsituation stehen nur wenige Indikatoren zur Verfügung, die sich einem Vergleich unterziehen lassen. Sehr ähnlich sind sich die Maße zur Reputation der Lehre (FOCUS) und der „Professorentipp" des CHE, in dem erfragt wurde, welche Universität der befragte Professor dem eigenen Kind zum Studium des von ihm vertretenen Faches empfehlen würde. Des Weiteren stehen in beiden Rankings Daten zur Betreuungssituation zur Verfügung. Während allerdings im FOCUS die zahlenmäßige Betreuungsrelation zwischen Studierenden und Lehrenden angegeben ist, stehen im CHE mit Ausnahme der Medizin nur Angaben der befragten Studierenden zur Zufriedenheit mit der Betreuungssituation zur Verfügung.[13] Ein Vergleich ist durchaus sinnvoll: Soll eine

[13] Der Verzicht auf die Angabe quantitativer Betreuungsrelationen ist dem CHE dabei nicht zur Last zu legen, sondern vielmehr als Versuch seriöser Datenerhebung zu honorieren.

günstige Betreuungsrelation überhaupt eine Aussagekraft haben, so müsste sie mit größerer Zufriedenheit mit der Betreuung einhergehen.

Die Ergebnisse (Tabelle 2) zeigen, dass auch bei der Einschätzung der Lehrqualität die Rankings zu unterschiedlichen Schlüssen kommen. Beim Vergleich von Betreuungsrelation und Beurteilung der Betreuung durch die Studierenden sind Korrelationswerte über 0,5 selten und werden bis auf die Fächer Bauingenieurwesen, Jura und Physik nur bei Weglassung von Ausreißern erreicht. Im Schnitt liegen die Korrelationen bei 0,23, was allenfalls einen mäßigen Zusammenhang darstellt. Keines der Fächer erreicht einen Wert von 0,7, der es erlauben würde, von einer „eben noch ausreichenden" Übereinstimmung zu sprechen. Auch im Fach Medizin, in dem auch das CHE-Ranking eine Betreuungsrelation ausweist, kann den Daten mit einem Korrelationskoeffizienten von 0,44 nur ein schwacher Zusammenhang attestiert werden – und dies auch nur bei Entfernung von Ausreißer-Werten.

Relativ gut stimmen die Reputationswerte von FOCUS und CHE überein. Sie liegen im Schnitt bei 0,57 und weichen damit deutlich ab von allen bisherigen Vergleichen. Lediglich im Falle der Psychologie wird eine Universität von beiden Rankings in entgegen gesetzte Extremgruppen einsortiert. Erklären lässt sich die relativ gute Übereinstimmung damit, dass die Datenquelle in beiden Rankings, nämlich die Urteile der befragten Professoren, identisch ist. Es ist anzunehmen, dass die Angaben der Professorinnen zur Reputation von Kollegen im Fach stabil sind, unabhängig davon, welche Institution die Daten erhebt. Trotz der Ähnlichkeit der Merkmale darf nicht in Vergessenheit geraten, dass in keinem Fall das CHE und der FOCUS zu übereinstimmender Ausweisung der Professorenreputation kommen. In allen Fächern wird mindestens ein Viertel aller Universitäten von den beiden Rankings nicht übereinstimmend gerankt.

Angesichts der Vielfältigkeit der Studiengänge an einem Fachbereich sind verlässliche Angaben zur Betreuungsrelation nur dann zu erhalten, wenn genau nach Studiengängen differenziert werden kann. Die Angaben des FOCUS zur Betreuungsrelation beruhen auf Auswertungen auf Fächergruppenebene (Systematik des Statistischen Bundesamtes: http://www.destatis.de/download/d/allg/stud_pruef.pdf; Zugriff am 02.06.2006); sie sollten daher mit größter Vorsicht betrachtet und nur als Tendenzen interpretiert werden.

Einzelindikatoren – Bewertung der Forschungssituation

Bei den Indikatoren für Forschungsleistungen sollte größere Übereinstimmung zu erwarten sein, da die hier verwendeten Indikatoren „objektive“ oder zumindest objektivierte Daten darstellen: Angaben zu eingeworbenen Drittmitteln oder zu Promotionen sollten unabhängig von individuell variierenden Bewertungen oder Interessen sein; obendrein werden sie vom FOCUS zum Teil aus amtlichen Statistiken übernommen. Bei dem Forschungsoutput, den wissenschaftlichen Veröffentlichungen, sind teilweise geringere Übereinstimmungen deshalb zu erwarten, weil der FOCUS ausschließlich Zitationen, das CHE-Ranking (das sich freilich häufiger als der FOCUS jeglicher Daten zu Publikationen enthält) teilweise nur Publikationen berichtet. Wir vergleichen dennoch Zitationen dort und Publikationen hier, sofern keine anderen Daten vorliegen, denn schließlich bleibt dann auch den Nutzern der Rankings, die sich ein Urteil über den Forschungsoutput verschaffen wollen, nichts anderes übrig, als die vorhandenen Daten heranzuziehen.

Bei den Drittmitteln liegen sechs von 18 Werten, also ein Drittel, nach Korrektur um Ausreißer noch ein Fall mehr im akzeptablen Bereich von $\geq$0,7; in vier Fällen liegt die Korrelation dagegen unter 0,4, was in zwei dieser Fälle auf Ausreißer zurückgeht, nach deren Ausschluss sich wenigstens Korrelationen von jeweils 0,56 ergeben. Insgesamt sind die Übereinstimmungen also überwiegend im mäßigen bis mittleren Bereich anzusiedeln, zumal der Wert 0,8 nur in zwei Fällen erreicht bzw. knapp überschritten wird. Besser sieht es mit den Promotionen aus: 12 der 18 Korrelationen betragen 0,7 oder mehr, in fünf (unter Einschluss der durch „positive“ Ausreißer allerdings unrealistisch günstig bewerteten Psychologie sogar sechs) Fällen liegt der Wert über 0,8. Studierende mit Ambitionen zur Promotion, die diese Angaben ihren Entscheidungen zu Grunde legen, werden jedenfalls nicht mit offensichtlich stark voneinander abweichenden Einschätzungen konfrontiert. Ob es Zufall ist, dass es ausgerechnet in drei sozialwissenschaftlichen Fächern – den beiden Wirtschaftswissenschaften und der Soziologie – Einschätzungen der Promotionsquoten eher Glückssache darstellen, bedürfte genauerer Analysen.

Tabelle 3: Indikatoren zur Forschung im Vergleich – CHE versus FOCUS

	Drittmittel		Promotionen		Output[(b)]	Forschungs-reputation
	Pearsons r[(a)]		Pearsons r[(a)]		Pearsons r	Kendalls Tau_b
Anglistik	.58	(.44)	.71		–	.43
Bauingenieurwesen	.81		.78		–	.74
Biologie	.41	(.63)	.78		.46 (Z)	.57
BWL	.57	(.77)	.28	(.57)	–	.42
Chemie	.33		.62		.60 (Z)	.68
Elektrotechnik	.72		.87		–	.70
Germanistik	.76		.80		–	.57
Geschichte	.70		.77		–	.63
Informatik	.52		.69		–	.60
Jura	–		.92		–	.64
Maschinenbau	.80		.87		–	.63
Mathematik	.76		.50		.19 (V)[(c)]	.42
Medizin	.63		.31	(.74)	.81 (Z)[(c)]	.62
Pädagogik	.55		.75		–.14 (V)[(d)]	.67
Physik	.48		.73		.74 (Z)	.61
Politikwissenschaft	.36	(.56)	.92		–	.66
Psychologie	.68		.82	(.67)	.21 (Z)	.64
Soziologie	.32	(.56)	.56	(.21)	–.26 (V)[(d)]	.71
VWL	.37		.36	(–.05)	.01 (V)	.29

– In (mindestens) einem Ranking keine Daten vorhanden

(a) Koeffizienten nach Elimination von Ausreißern in Klammern

(b) Z: Es werden aus beiden Rankings Zitationsdaten verglichen
V: Veröffentlichungsdaten (CHE) werden mit Zitationsdaten (FOCUS) verglichen

(c) CHE-Daten aus dem Jahr 2006

(d) Kendalls Tau_b

Wesentlich häufiger besteht das Risiko einer unzuverlässigen Einschätzung wiederum, wenn man sich an den Publikationen orientiert.[14] Gewiss treten besonders schlechte Übereinstimmungswerte vor allem dann auf,

[14] Weil die unserer Auswertung zugrunde liegenden CHE-Rankings der Jahre 2003 bis 2005 relativ wenig Daten zu Publikationen bzw. Zitationen auswiesen, haben wir aus dem zum Zeitpunkt der Abfassung dieses Textes gerade erschienenen CHE-Ranking 2006 noch Daten zu zwei Fächern erfasst, zu denen hier erstmals Publikationsdaten vorgelegt wurden.

wenn ein Ranking nur Publikationen, das andere nur Zitationen ausweist – wobei sich die Frage nach dem Wert jedes einzelnen dieser Indikatoren stellt, wenn das, was Wissenschaftler produzieren, offenbar wenig mit der durch Zitationen belegten Rezeption der Produktion zu tun hat. Aber auch wenn beide Rankings Zitationen ausweisen, liegen nur in zwei von fünf Fällen zufriedenstellende Werte vor.

Einzig bei der Forschungsreputation besteht höhere Übereinstimmung. Ähnlich wie im Falle der Lehrreputation ist der durchschnittliche Tau_b-Wert mit 0,59 vergleichsweise hoch; in zwei Fächern (Bauingenieurwesen und Soziologie) wird der Wert 0,7 überschritten, der auf eben noch ausreichende Übereinstimmung zwischen den Rankings hinweist, in weiteren vier Fächern liegt Tau_b nahe an diesem Wert. Die beiden Wirtschaftswissenschaften und die Anglistik weichen mit Werten zwischen 0,3 und 0,4 von den ansonsten relativ hohen Werten um 0,6 deutlich ab.

Fazit

Die vergleichende Betrachtung der drei prominentesten deutschen Hochschulrankings zeigt: Die Rankings erreichen nur einen unbefriedigenden Grad an Übereinstimmung, die Angaben zu Globalurteilen sind kaum verlässlich, lediglich die Reputation in Lehre und Forschung sowie die Daten zu den Promotionsquoten weisen einen zufriedenstellenden Grad an Übereinstimmung auf.

Die Ergebnisse machen deutlich: Die in den Medien veröffentlichen Rankings sind kein adäquates Mittel zur Orientierung, da ihre Resultate nicht übereinstimmen und sie damit Universitäten nicht zuverlässig bewerten. Sie messen zwar und geben sich Mühe, dabei sehr exakt zu wirken. Was genau da aber eigentlich gemessen wird, ist nur selten klar. Zu sehr verschwimmen in Folge von Stichprobenfehlern, Aggregierung und Gruppenbildung die eigentlichen Merkmale. Sicher wäre – nimmt man den Orientierungsanspruch ernst – künftigen Studierenden mehr geholfen, wenn sie nicht-wertende Studienführer zur Hand hätten, die Universitäten und Fachbereiche qualitativ beschreiben, anstatt sie auf Basis pseudo-objektiver Zahlen zu bewerten.

Hochschulrankings werden nicht von der Bildfläche verschwinden, auch wenn unsere Ergebnisse zeigen, dass die Rankings ihrem Anspruch, künftigen Studierenden Orientierung zu verschaffen, nicht gerecht wer-

den können. Ganz im Gegenteil befinden sich die Rankings im Aufschwung, wie es z.B. die Bemühungen des CHE zeigen, auf internationaler Ebene Qualitätskriterien für Hochschulrankings zu kanonisieren.[15] Die sich entfaltende „Ranking-Kultur" als Begleiterscheinung der wettbewerblichen Hochschuldifferenzierung wird mit sich bringen, dass sich die Signalwirkung des Hochschulortes für Arbeitgeber erhöht. Die Chancen der Absolventinnen werden mit der von diesem oder jenem Ranking induzierten Wahrnehmung von Universitäten stehen und fallen, nicht aber mit tatsächlichen Qualitätsunterschieden der Ausbildung.

Diese mögliche Entwicklung stößt uns auf die Frage nach Verantwortlichkeit: Interessant ist im medial organisierten Ranking-Betrieb, dass von Institutionen gerankt wird, die in keiner Weise verantwortlich gemacht werden können für Fehlbewertungen. Solange für Studienanfänger das Bild, das Rankings von der „Wunsch-Uni" verschafft haben, in Einklang ist mit den Verhältnissen, die sie an der Uni ihrer Wahl vorfinden, ist dies unproblematisch. Stellt man hier ein Missverhältnis zwischen auf Basis von Rankings erwarteten und tatsächlich vorgefundenen Studienbedingungen fest, bleibt der Schaden, der durch die Korrektur der Entscheidung entsteht, bei den Studierenden. Ein von Rankings ausgesprochenes Qualitätsversprechen ist allerdings an keiner Stelle einklagbar – weder für die Studierenden noch für ihre künftigen Arbeitgeberinnen und Arbeitgeber.

Literatur

Bortz, Jürgen/Döring, Nicola 2002: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 3., überarbeitete Auflage. Berlin, Heidelberg, New York: Springer

Brenner, Hermann/Kliebsch, Ulrike 1996: Dependence of Weighted Kappa on the Number of Categories. In: Epidemiology, Vol. 7, Heft 2, S. 199-202

Guggenmoos-Holzmann, Irene 2005: How Reliable Are Chance-corrected Measures of Agreement? In: Statistics in Medicine, Vol. 12, S. 2191-2205

[15] Das CHE hat sich mit dem CEPES (Centre européen pour l'enseignement supérieur, Organisation der UNESCO) und dem in Washington ansässigen IHEP (Institute for Higher Education Policy, finanziert von us-amerikanischen, privaten Stiftungen) zur International Ranking Expert Group (IREG) zusammengeschlossen, die auf ihrem jüngsten Treffen im Mai 2006 in Berlin 16 Prinzipien für die Bewertung von Hochschulen verabschiedete (http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf; Zugriff am 01.06.2006).

Liebeskind, Uta/Ludwig-Mayerhofer, Wolfgang 2005: Auf der Suche nach der Wunsch-Universität – im Stich gelassen. Anspruch und Wirklichkeit von Hochschulrankings. In: Soziologie, Vol. 34, Heft 4, S. 442-462

Lienert, Gustav A./Raatz, Ulrich 1994: Testaufbau und Testanalyse. 5. Auflage. Weinheim: Beltz, PsychologieVerlagsUnion

Mächtle, Tomas/Witthaus, Udo 2002: Bildungstests – mehr Transparenz für Bildungsinteressierte? Eine Einschätzung zum Nutzen von Tests für Weiterbildungsinteressierte. In: Balli, Christel et al. (Hg.): Qualitätsentwicklung in der Weiterbildung. Zum Stand der Anwendung von Qualitätssicherungs- und Qualitätsmanagementsystemen bei Weiterbildungsanbietern. Bonn: Diskussionspapiere des Bundesinstitutes für Berufsbildung. S. 45-84

Meinefeld, Werner 2000: Hochschulranking. Eine unsichere Basis für Entscheidungen. In: Forschung & Lehre, Heft 1, S. 26-29.

Pechar, Hans 1997: Leistungstransparenz oder Wünschelrute? Über das Ranking von Hochschulen in den USA und im deutschsprachigen Raum. In: Altrichter, Herbert et al. (Hg.): Hochschulen auf dem Prüfstand. Innsbruck: Studienverlag. S. 157-178

Uebersax, John S. 1987: Diversity of Decision-Making Models and the Measurement of Interrater Agreement. In: Psychological Bulletin, Vol. 101, Heft 1, S. 140-146

Wirtz, Markus/Caspar, Franz 2004.: Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen, Bern, Toronto, Seattle: Hogrefe

Wissenschaftsrat 2004: Empfehlungen zu Rankings im Wissenschaftssystem. Teil 1: Forschung. http://www.wissenschaftsrat.de/texte/6285-04.pdf (Zugriff am 03.06.2006)