

# Hochschul-Rankings: Probleme, Lösungsvorschläge und ein Modell auf Basis des Studentenpisa-Tests

Rüdiger Mutz, Hans-Dieter Daniel (ETH Zürich)

## Zusammenfassung:

*Der SPIEGEL hat 1989 das erste Hochschulranking in Deutschland veröffentlicht. Seitdem sind zahlreiche nationale und internationale Rankings entstanden, die große Beachtung gefunden haben. Allerdings weisen solche Rangordnungen, wenn sie auf Befragungen basieren, eine Reihe von methodischen Problemen auf (Stichprobenfehler, Konsequenzen der Bildung von Gesamtindizes, Messfehler und urteilsverzerrende Faktoren). Der Beitrag stellt diese Probleme dar und skizziert Lösungen am Beispiel des Studentenpisa-Tests des SPIEGEL. Ziel ist es, ein auf studentischen Befragungen basierendes Hochschul-Ranking zu entwickeln, das die in der Fachliteratur am häufigsten genannten methodischen Probleme überwindet. Das vorgeschlagene Verfahren wird anhand des Studienfachs Wirtschaftswissenschaften illustriert: Für dieses Fach wird ein Ranking der deutschen Universitäten vorgelegt.*

## 1. Einleitung

In den letzten zwanzig Jahren sind in der ganzen Welt Rankings von Hochschulen entstanden (Kroth & Daniel, 2008; Usher & Savino, 2007). Erstmals wurde in Deutschland im Jahr 1989 ein Hochschul-Ranking vom Nachrichtenmagazin DER SPIEGEL veröffentlicht (Daniel & Hornbostel, 1993). Unter Rankings werden allgemein Rangordnungen von Hochschulen entsprechend einer Qualitätsdimension verstanden. Diese Dimension ergibt sich aus unterschiedlich gewichteten Indikatoren, die zu einem Gesamtindex, zum Beispiel für die Qualität der Hochschule, zusammengefasst werden. Rankings haben eine immer größere Bedeutung als Informationsquelle: Erstens für Studierende bei der Wahl einer Hochschule, zweitens für das Marketing der Hochschulen und drittens für Bildungsministerien als Grundlage für die leistungsbezogene Mittelvergabe. Das Aufkommen von Rankings wird unter anderem auf die steigenden Studierendenzahlen, die stärkere Internationalisierung der Hochschulen und den insgesamt verschärften Wettbewerb unter den Hochschulen zurückgeführt (van der Wende, 2008; Wissenschaftsrat, 2008). Dieser wachsenden gesellschaftlichen Bedeutung wird seitens sozialwissenschaftlicher, insbesondere psychologischer Forschung in Deutschland im Gegensatz zu Großbritannien und den USA nicht in gleichem Maße entsprochen (Engel & Krekler, 2001; Leberherz et al., 2005; Mutz & Daniel, 2007). Ziel des Beitrags ist es, am Beispiel des Studentenpisa-Tests des SPIEGEL ein befragungsbasiertes Hochschul-Ranking-Verfahren zu entwickeln, das für die in der Fachliteratur am häufigsten genannten methodischen Probleme Lösungsansätze aufzeigt und das internationalen Standards genügt.

## 2. Rankings im Spiegel der Kritik

Hinweise auf methodische Probleme von Rankings finden sich in zahlreichen Übersichtsarbeiten (z. B. Dill & Soo, 2005; Hornbostel, 2007; Kroth & Daniel, 2008). Es lassen sich im Wesentlichen vier zentrale methodische Kritikpunkte an Hochschul-Rankings anführen: 1) Stichprobenfehler, 2) Bildung von Gesamtindizes, 3) Messfehler und 4) urteilsverzerrende Faktoren.

### 2.1 Stichprobenfehler

Bei befragungsbasierten Hochschul-Rankings entstehen Probleme im Zusammenhang mit der Stichprobenziehung, sogenannte *Stichprobenfehler* (Lohr, 2008). Die Idee der Stichprobenziehung ist es, anhand einer kleinen Auswahl von Personen aus einer Grundgesamtheit (z. B. 30.000 Studierende einer Hochschule) eine zuverlässige Aussage über die Grundgesamtheit zu machen (Cochran, 1977). Um einen zuverlässigen Schluss von der Stichprobe auf die Grundgesamtheit zu ziehen, ist es erforderlich, dass diese Auswahl per Zufall getroffen wird („repräsentative Stichprobe“): Jede Person der Grundgesamtheit muss die gleiche oder eine exakt bestimmbare Wahrscheinlichkeit haben, Teil der Stichprobe zu werden. Die meisten Befragungen, die als Informationsquelle für Rankings dienen, basieren jedoch auf einer nicht-repräsentativen Selbstselektion der Befragten: Studierende beispielsweise wählen selber aus, ob sie an einer Befragung teilnehmen wollen oder nicht. Diejenigen, die sich an der Befragung beteiligen, unterscheiden sich hinsichtlich relevanter Merkmale in der Regel von denjenigen, die sich nicht beteiligen. Selbst wenn wie in den PISA-Studien (Programme for International Student Assessment) der OECD (Adams & Wu, 2002) Zufallsstichproben gezogen werden, können die ausgewählten Personen sich weigern, an der Befragung teilzunehmen. Nicht antwortende Personen könnten anders antworten als teilnehmende Personen, was als „Nonresponse Error“ bezeichnet wird (Lynn, 2008). Hinzukommt, dass Befragungen häufig wie beim Studententpisa-Test online erfolgen, was einen Personenkreis systematisch ausschließt, der keinen eigenen Computer besitzt bzw. keinen Zugang zum Internet hat. Die Verallgemeinerbarkeit der Ergebnisse ist damit mehr oder weniger stark eingeschränkt.

### 2.2 Bildung von Gesamtindizes

Ein weiterer zentraler Kritikpunkt an Hochschul-Rankings ist die Bildung von *Gesamtindizes*. Nach Berghoff und Federkeil (2009, S. 42) werden Universitäten häufig auf einer Qualitätsdimension verglichen. Diese Dimension ergibt sich aus unterschiedlich gewichteten Indikatoren, z. B. der Anzahl von Nobelpreisträgern, die an der Hochschule forschen, der Zufriedenheit von Studierenden mit der Studiensituation an ihrer Hochschule und dem Ruf der Hochschule bei Arbeitgebern. Diese Indikatoren werden zu einem Gesamtindex, zum Beispiel für die Qualität einer Hochschule, zusammengefasst. Die „Qualität“ der einzelnen Hochschulen wird dann in Form einer Liga-Tabelle dargestellt. Mit der Bildung eines Gesamtindex sind jedoch zwei Probleme verbunden: Erstens impliziert ein Gesamtindex eine einheitliche zugrundeliegende Qualitätsdimension, auf der Hochschulen verglichen werden können. Häufig werden jedoch sehr unterschiedliche Indikatoren zu einer Qualitätsdimension zusammengefasst. Usher und Savino (2007) klassifizieren auf der Grundlage einer Erhebung von weltweiten Hochschul-Rankings die gefundenen Einzelindikatoren in vier Gruppen, die als unterschiedliche Qualitätsdimensionen aufgefasst werden können: (a) Merkmale und Fähigkeiten der Studi-

enanfängerinnen und -anfänger (personelle Input-Indikatoren), (b) finanzielle und materielle Ressourcen für Studierende und Lehrende (materielle Input-Indikatoren), (c) Fähigkeitsprofile von Hochschulabsolventinnen und -absolventen (Output-Indikatoren) und (d) Erwerbsquote und Einkommen der Hochschulabsolventinnen und -absolventen, Forschungsleistung der Wissenschaftlerinnen und Wissenschaftler, Reputation der Hochschule (Outcome-Indikatoren).

Zweitens ist die Gewichtung der Indikatoren im Gesamtindex willkürlich und variiert sehr stark zwischen den einzelnen Hochschul-Rankings (Usher & Savino, 2007). Darüber hinaus ändern sich die Gewichtungen über die Jahre, sodass Vergleiche über die Jahre problematisch werden. In dem von der britischen Zeitung Times Higher Education Supplement veröffentlichten World University Ranking (THES) für das Jahr 2009 geht beispielsweise die Exzellenz der Forschung mit 20% in den Gesamtindex ein ([www.timeshighereducation.co.uk](http://www.timeshighereducation.co.uk)). Dagegen geht in das Hochschul-Ranking, das von der Shanghai Jiao Tong Universität im selben Jahr publiziert wurde ([www.arwu.org](http://www.arwu.org)), die Forschung mit 90% ein. Das Centrum für Hochschulentwicklung (CHE) in Deutschland ist daher von der Idee eines gewichteten Gesamtindex ganz abgerückt und konzipiert seine „Rankings“ mehrdimensional. Auf Rangordnungen von Universitäten wird ganz verzichtet. Stattdessen wird zwischen einer Spitzengruppe (grün), einer mittleren Gruppe (gelb) und einer unterdurchschnittlichen Gruppe (rot) farblich unterschieden.

### 2.3 Messfehler

Ein weiterer zentraler Kritikpunkt betrifft die Messung der Qualitätsdimension bzw. der Qualitätsindikatoren und der damit verbundenen *Messfehler*. So ist die Messung an sich in den Sozialwissenschaften schon ein Problem: Vergleichbar der Heisenbergschen Unschärferelation (Heisenberg, 1927) kann die Messung selbst das zu Messende verändern, was sich in Goodhart's Law ausdrückt (Goodhart, 1981; Elton, 2004): Wenn ein Qualitätsindikator selbst zum Ziel wird, dann hört er auf, ein gutes Maß zu sein. Hochschulen werden beispielsweise versuchen, durch unterschiedliche Strategien (z. B. weniger strenge Notengebung) hohe Werte auf den Qualitätsindikatoren (z. B. hohe Absolventenquote) zu erreichen und verfälschen damit den Indikator. Darüber hinaus erscheinen nach Salmi und Saroyan (2007, S. 8) Unterschiede zwischen Institutionen in den Rangtabellen größer als sie tatsächlich sind. Kleinste Mittelwertunterschiede zwischen Hochschulen können schon zu Rangunterschieden in der Liga-Tabelle führen und die Illusion bedeutsamer Unterschiede hervorrufen. So wird häufig eine Verbesserung in der Tabelle um ein bis zwei Rangplätze im Vergleich zum Vorjahr von den Hochschulen als großer Erfolg verbucht, obwohl die bessere Platzierung meist keine statistisch bedeutsamen Unterschiede widerspiegelt. Dies liegt zum einen darin, dass die zugrundeliegende Qualitätsdimension nicht ganz exakt erfasst werden kann. Deshalb kann nicht trennscharf zwischen Hochschulen, die in der Liga-Tabelle nahe beieinanderliegen, unterschieden werden („Messfehler“). Zum anderen besteht das Problem darin, dass Qualitätsunterschiede zwischen Hochschulen („Signal“) viel kleiner sind als Unterschiede innerhalb der Hochschulen („Rauschen“), was sich auf die statistische Signifikanz als Signal-Rausch-Abstand auswirkt. Aus der Untersuchung zur Beurteilung der Qualität von Lehrveranstaltungen wissen wir beispielsweise, dass lediglich 20 bis 30% der Variabilität in den studentischen Bewertungen auf Unterschiede zwischen Lehrveranstaltungen und 70 bis 80% auf Unterschiede in den studentischen Bewertungen innerhalb der Veranstaltungen zurückgehen (Rindermann, 2001). Daher verzichtet beispielsweise das CHE ganz auf Liga-Tabellen und weist stattdessen Rang-Gruppen aus. Dies

ist jedoch auch keine Lösung für das Messproblem. Die mittlere Gruppe kann nicht eindeutig als „mittelmäßige“ Gruppe interpretiert werden, da die Anordnung in der Mitte statistisch auch Folge einer zu kleinen Stichprobe sein kann. Eine zu hoher Streuung der studentischen Beurteilungen innerhalb eines Studienfaches kann ebenfalls zu einer Platzierung in der Mitte führen. Die Platzierung im Mittelfeld eines Rankings ist häufig nicht Ausdruck mittelmäßiger Leistung, sondern Ausdruck der statistischen Unsicherheit des ermittelten Ergebnisses.

#### 2.4 Urteilsverzerrende Faktoren

Darüber hinaus wird kritisiert, dass Rankings dem *Einfluss von urteilsverzerrenden Faktoren* (sogenannte *Bias-Faktoren*) unterliegen, die von den Hochschulen selbst nicht beeinflussbar sind (Marsh, 2007; Spiel, 2001). So gelten zum Beispiel für die studentischen Bewertungen von Lehrveranstaltungen oder für Wissenstests das Geschlecht, das Alter, das Interesse am Thema oder die beruflichen Vorerfahrungen der Studierenden als mögliche Bias-Faktoren. Würden beispielsweise Frauen unabhängig von der Hochschule, an der sie studieren, bessere Testergebnisse als Männer erzielen, so würden Hochschulen mit höherem Frauenanteil (z. B. Pädagogische Hochschulen) besser abschneiden als Hochschulen mit einem geringeren Frauenanteil (z. B. Technische Hochschulen). Dies wäre ein unfairer Hochschulvergleich.

### 3. Vorschlag für ein verbessertes Ranking-Verfahren

Vor dem Hintergrund der oben genannten Probleme muss ein Hochschul-Ranking-Verfahren Lösungsansätze für diese vier zentralen Problembereiche formulieren: (1) Wie kann die Gruppe der Befragten bestimmt werden, sodass sie die Populationsverhältnisse mit möglichst geringem Fehler widerspiegelt („*Stichprobenfehler*“)? (2) Wie können ein- oder mehrdimensionale Skalen entwickelt werden („*Gesamtindizes*“)? (3) Wie können Hochschulen unter Berücksichtigung des Messfehlers in eine Rangordnung gebracht werden („*Messfehler*“) und (4) Wie können die Rangordnungen für einen fairen Hochschulvergleich im Hinblick auf mögliche Bias-Faktoren korrigiert werden („*urteilsverzerrende Faktoren*“)?

Der folgende Vorschlag eines verbesserten Ranking-Verfahrens beruht im Wesentlichen auf einem Ansatz von Mutz und Daniel (2007), auf methodischen Konzepten, die im Rahmen des von der OECD initiierten PISA-Projektes entwickelt wurden (Adams & Wu, 2002), und auf weiteren Beiträgen von ausländischen Bildungsforschern (Goldstein & Leckie, 2008; Goldstein & Spiegelhalter 1996; Lubrano, 2009). Es lassen sich folgende vier Lösungen zu den oben genannten Problemen formulieren:

#### 3.1 Lösung zum Problem des Stichprobenfehlers

Ein aussagekräftiges Hochschul-Ranking setzt die Ziehung von Zufallsstichproben voraus („repräsentative Stichprobe“): So könnten beispielsweise auf der Grundlage von Adressen, die den Hochschulen zu Verfügung stehen, Studierende zufällig ausgewählt werden. Ein anderes, weniger aufwendiges Verfahren besteht darin, zunächst möglichst viele Studierende mit einer Befragung zu erreichen. Nach der Befragung wird dann versucht, durch eine sogenannte „*Post-stratifizierende Adjustierung*“ die Stichprobe der Befragten durch eine Form von Gewichtung der Grundgesamtheit im Hinblick auf wichtige soziodemografische Merkma-

le anzugleichen (Biemer & Christ, 2008). Hierfür sind Informationen über die Population der Befragten (z. B. Altersverteilung, Frauenanteil) notwendig. Diese Informationen können vom Statistischen Bundesamt bezogen werden. Sind beispielsweise in der Stichprobe 50 Frauen im Alter von 20 Jahren vertreten und umfasst die Population 500 Frauen im Alter von 20 Jahren, so muss jede Frau in der Stichprobe im Alter von 20 Jahren mit dem Faktor 10 gewichtet werden, um die Populationsverhältnisse korrekt widerzuspiegeln.

### 3.2 Lösung zum Problem der Bildung von Gesamtindizes

Anstelle eines Gesamtindex wird ein mehrdimensionales Ranking favorisiert. So können im Studententpisa-Test die einzelnen Bereiche wie Politik, Wirtschaft oder Geschichte beispielsweise unterschiedliche Dimensionen widerspiegeln. Es können aber auch unterschiedliche Aufgaben innerhalb eines Bereichs (z. B. Wirtschaft) unterschiedliche Dimensionen abbilden (Betriebswirtschaft, Volkswirtschaft, ...). Um die Anzahl Dimensionen zu bestimmen und Messwerte für jede Person auf diesen Dimensionen zu bestimmen, eignet sich sehr gut das sogenannte Rasch-Modell (Ayala, 2009). Nach dem Rasch-Modell ist die Wahrscheinlichkeit, eine Aufgabe z. B. des Studententpisa-Tests zu lösen,  $P(X=1)$ , von zwei Faktoren abhängig: einerseits von der Fähigkeit oder dem Wissensstand der befragten Personen ( $\theta$ ), andererseits von der Schwierigkeit der Aufgabe  $j$  ( $\sigma_j$ ). Leichte Aufgaben können demnach von Personen mit geringem Kenntnissniveau gelöst werden, schwierige Aufgaben nur von Personen mit hohem Fähigkeitsniveau ( $e=e$ -Funktion 2.71\*):

$$P(X = 1 | \theta, \sigma_j) = \frac{e^{(\theta - \sigma_j)}}{1 + e^{(\theta - \sigma_j)}}$$

Mit dem Rasch-Modell verbinden sich eine Reihe von Vorteilen: Eine zentrale Eigenschaft ist die Eindimensionalität. Passt das Rasch-Modell auf die Daten, so liegt den betreffenden Aufgaben eine eindimensionale latente Wissensdimension (z. B. Allgemeinwissen in Wirtschaft) zugrunde. Im anderen Falle müssen mehrdimensionale Rasch-Modelle geschätzt werden. Ein weiterer Vorteil des Rasch-Modells liegt darin, dass die Aufgabenparameter unabhängig von der gewählten Personenstichprobe und die Personenparameter unabhängig von der gewählten Stichprobe von Aufgaben geschätzt werden können („spezifische Objektivität“). Der Vergleich von zwei Personen ist demnach unabhängig von der spezifischen Auswahl von Aufgaben. Diese Eigenschaft wird auch im Studententpisa-Test genutzt, in dem nicht alle Befragten alle Aufgaben zu einem Wissensbereich (36 Items) lösen müssen, sondern jeweils nur 9 Aufgaben, um den Kenntnisstand in diesem Wissensbereich zu erfassen. Dennoch lassen sich in Rahmen eines „Equating“-Verfahrens die Personen, die unterschiedliche Testteile gelöst haben, miteinander vergleichen (Holland, Dorans & Petersen, 2007).

Ergänzend sei darauf hingewiesen, dass für einen Vergleich von Hochschulen sich weniger die aus dem Rasch-Modell geschätzten Personenparameter eignen als vielmehr sogenannte „plausible Werte“. Diese werden unter Annahme einer Wahrscheinlichkeitsverteilung der Personenparameter (z. B. Normalverteilung) und durch Berücksichtigung von weiteren Deter-

minanten (Alter, Geschlecht, ...) im Rahmen eines alternativen Schätzansatzes in der Statistik (Bayes-Statistik, a-posteriori-Verteilung) bestimmt (Adams & Wu, 2002).

### 3.3 Lösung zum Problem des Messfehlers

Die Ableitung einer Rangordnung von Hochschulen erfolgt über das statistische Verfahren der Mehrebenenanalyse (Hox, 2002; Snijders & Bosker, 1999), das auf die „plausiblen Werte“ angewendet wird (siehe Kasten *Mehrebenenanalyse*). Es werden Mehrebenenanalysen für jedes Set von plausiblen Werten berechnet (Adams & Wu, 2002). Im letzten Schritt werden Empirische Bayes-Schätzer für die Mittelwerte oder mittleren Ränge auf der Qualitätsdimension berechnet und über das Set von plausiblen Werten gemittelt (siehe Kasten *Empirische Bayes-Schätzer*). Werden diese Empirischen Bayes-Schätzer je Hochschule auf der Wissensdimension ranggeordnet, so ergibt sich das gewünschte Hochschul-Ranking. Darüber hinaus lässt sich für jeden Schätzwert einer Hochschule ein Fehlerbereich (Vertrauensintervall) bestimmen, innerhalb dessen der wahre Populationswert zu finden ist.

#### **Mehrebenenanalyse:**

Die Mehrebenenanalyse ist ein statistisches Analyseverfahren, das die hierarchische Struktur der Daten, wie sie in Rankings anfallen, berücksichtigt. So sind – vergleichbar einer russischen Puppe – Studierende in Studiengänge und Studiengänge in Hochschulen gruppiert (Hox, 2002). Es ist davon auszugehen, dass innerhalb einer Aggregateinheit (z. B. Studiengang) die Personen ähnlicher sind hinsichtlich ihrer Beurteilungen oder ihrem Wissen als zwischen unterschiedlichen Aggregateinheiten. In der Statistik wird hierfür der Begriff „Abhängigkeit“ oder „Intraklassen-Korrelation“ gebraucht (Hox, 2002, S.5f). Die statistischen Tests setzen jedoch unabhängige Messungen voraus. Liegt Abhängigkeit der Messungen vor, so ist der Standardfehler eines Parameters zu klein, viele Ergebnisse werden fadenscheinig „statistisch signifikant“. Mit dem Konzept des Standardfehlers wird der Tatsache Rechnung getragen, dass mit einer Stichprobe nur mit einem gewissen Fehler der wahre Parameter in der Population geschätzt werden kann. Er wird stark von der Stichprobengröße bestimmt: Je größer die Stichprobe, desto kleiner der Standardfehler, desto genauer wird die Schätzung des wahren Populationsparameters. Im Falle von Abhängigkeiten sinkt die Stichprobengröße und steigt damit der Standardfehler.

#### **Empirische Bayes-Schätzer:**

Im Rahmen der Mehrebenenanalyse ist es möglich Empirische Bayes-Schätzer zu berechnen (Snijders & Bosker, 1999, S. 58; Hox, 2002, p. 28f). Diese Schätzer ergeben sich durch eine Schrumpfung des Rohmittelwerts einer Hochschule auf der Qualitätsdimension in Richtung des Gesamtmittelwerts über alle Hochschulen. Diese Schrumpfung ist umso stärker, je kleiner die Stichprobe an Befragten an dieser Hochschule ist und je mehr diese Hochschule in ihrem Rohmittelwert vom Gesamtmittelwert abweicht (Reliabilität des Parameters). Diese letztlich „verzerrten“ Schätzer liefern statistisch gesehen jedoch bessere Schätzungen der Populationskennwerte als die Rohmittelwerte selber.

### 3.4 Lösung zum Problem der urteilsverzerrenden Faktoren

Wie im Problemaufriss zu Hochschul-Rankings ausgeführt, können Faktoren Einfluss auf die Ergebnisse des Rankings nehmen, die von den Hochschulen selbst nicht kontrollierbar sind. Die Bereinigung der Rankings um den Einfluss von Bias-Faktoren erfolgt dabei auf statistischem Wege über eine sogenannte Kovarianzadjustierung (Rubin, 1973). Bei der Kovarianzadjustierung werden die (Residual-) Werte für jeden Studierenden bestimmt. Das ist der Anteil der Rohwerte, der nicht durch die Bias-Faktoren in einer Regression vorhergesagt werden kann. Nach der Adjustierung kann ein Ranking so interpretiert werden, als ob alle Studierende die gleiche Ausprägung in allen berücksichtigten Bias-Faktoren haben, d. h. der Einfluss von Bias-Faktoren auf die Platzierung einer Hochschule im Ranking wird auf statistischem Wege ausgeschlossen.

## 4. Ein Hochschul-Ranking am Beispiel des Studententpisa-Tests

Das vorgestellte Verfahren soll im Folgenden am Beispiel des Studententpisa-Tests illustriert werden. Ziel ist es, ein Ranking von Universitäten im Hinblick auf das Allgemeinwissen in Wirtschaft am Beispiel von Studierenden des Studienfachs „Wirtschaftswissenschaften“ zu erstellen. Im Sinne der Klassifikation der Indikatoren von Usher und Savino (2007) handelt es sich dabei um ein Ranking im Hinblick auf den studentischen Input einer Hochschule.

### 4.1 Stichprobenfehler

Für die Auswahl von Hochschulen und Studierenden wurden folgende Kriterien verwendet: Es wurden aus dem Gesamtdatensatz des Studententpisa-Test Studierende ausgewählt, die im Fragebogen angeben haben, das Studienfach „Wirtschaftswissenschaften“ im Hauptfach an einer Universität (nicht Fachhochschule) in Deutschland zu studieren. Diese Universitäten müssen entsprechend der Studierendenstatistik des Statistischen Bundesamts „Wirtschaftswissenschaften“ als Studienfach im Wintersemester 2008/2009 angeboten haben (Statistisches Bundesamt, 2009). Zusätzlich müssen mindestens zehn Studierende einer Hochschule am Studententpisa-Test teilgenommen haben, um aussagefähige Ergebnisse zu ermöglichen. Mit diesen Kriterien erhält man eine Stichprobe von 5862 Studierenden aus 57 Universitäten. Es zeigten sich statistisch bedeutsame Unterschiede zwischen der Stichprobe und der Population, zum einen im Frauen- und Männeranteil, zum anderen in den Größenverhältnissen der Universitäten: So nahmen an Universitäten mit hoher Anzahl Studierender in „Wirtschaftswissenschaften“ prozentual weniger an der Untersuchung teil als an Universitäten mit geringerer Anzahl Studierender. Daher wurden die Analysen im Sinne der „Post-stratifizierenden Adjustierung“ gewichtet.

### 4.2 Bildung von Gesamtindizes

Im zweiten Schritt wurde mittels eines Rasch-Modells geprüft, ob die Aufgaben im Bereich „Wirtschaft“ eindimensional oder mehrdimensional aufgebaut sind. Es zeigte sich, dass ein modifiziertes Rasch-Modell (sogenanntes 2PL-Modell) mit der Annahme unterschiedlicher Trennschärfe der Aufgaben sehr gut mit den Daten übereinstimmt. Die Trennschärfe gibt an, in welchem Maße es eine Aufgabe erlaubt, zwischen Studierenden mit unterschiedlich hohem

Wissen zu trennen. Das Modell zeigt einen Passungsgrad von 99% (Bentler's Comparative Fit Index). Es liegt damit in hohem Maße Eindimensionalität der Wissensdimension („Allgemeinwissen in Wirtschaft“) trotz der unterschiedlichen Trennschärfe der Aufgaben vor. Die Erstellung von multidimensionalen Hochschul-Rankings sind hier nicht erforderlich. Somit ist der Wissenstest sehr gut für ein Hochschul-Ranking geeignet.

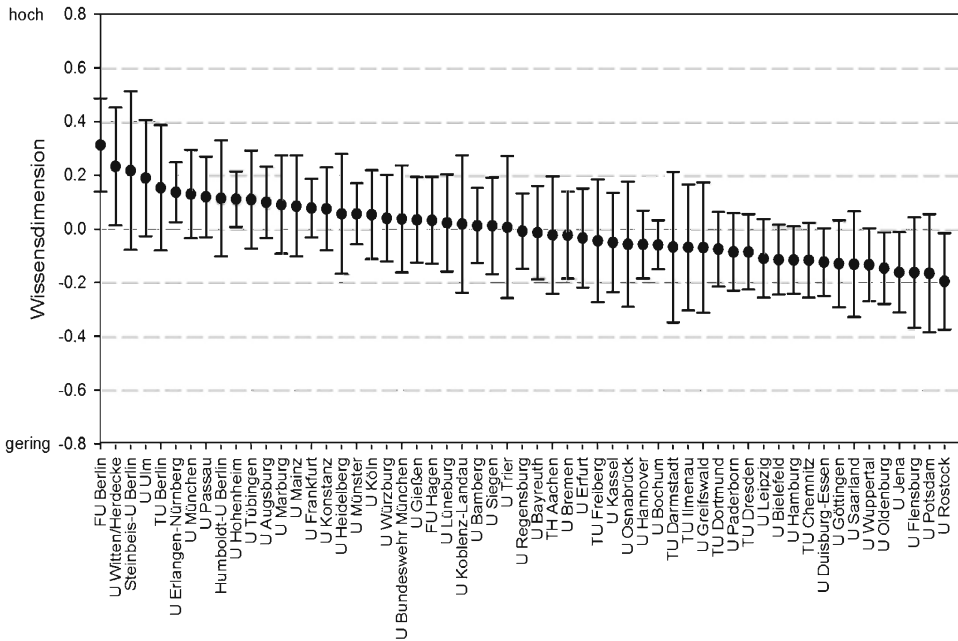
### 4.3 Messfehler

Im Weiteren ging es darum, mittels Mehrebenenanalyse unverzerrte Mittelwerte für jede Hochschule auf der Wissensdimension zu berechnen und die dazugehörigen Standardfehler zu bestimmen. Diese Mittelwerte – in eine Rangordnung gebracht – ergeben das Ranking der Universitäten. Die Mehrebenenanalyse wurde auf der Grundlage der aus dem ersten Analyseschritt geschätzten „plausiblen Werte“ berechnet. Die Ergebnisse der Analyse zeigen, dass die Unterschiede zwischen den Universitäten auf der Wissensdimension zwar statistisch signifikant von Null abweichen, aber sehr klein sind: Die Mittelwertunterschiede der Hochschulen machen rund 4% an der Gesamtvariabilität im Allgemeinwissen in Wirtschaft über alle Studierenden der Stichprobe aus. Die verbleibenden 96% der Gesamtvariabilität gehen auf Wissensunterschiede der Studierenden innerhalb der Universitäten zurück. Einzelne Hochschulen lassen sich daher mit dem Wissenstest nicht zuverlässig trennen, allenfalls Gruppen von Hochschulen.

### 4.4 Urteilsverzerrende Faktoren

Im letzten Analyseschritt wurden weitere mögliche urteilsverzerrende Faktoren in die Mehrebenenanalyse einbezogen, um ein adjustiertes Hochschul-Ranking zu erstellen: Es zeigt sich, dass männliche Studierende mit höherem Lebensalter und Semesterzahl über ein höheres Allgemeinwissen in „Wirtschaft“ verfügen als die übrigen Studierenden. Das Allgemeinwissen in Wirtschaft ist auch umso höher, je öfter das Nachrichtenmagazin „DER SPIEGEL“ (weniger das Magazin „STERN“) gelesen und SPIEGEL ONLINE genutzt wird. Einen starken Einfluss auf die Rangordnung der Hochschulen haben diese Faktoren allerdings nicht. So weist das (hier nicht gezeigte) unkorrigierte Hochschul-Ranking einen hohen Zusammenhang (Kendall's Tau-b von 0.69) mit dem adjustierten Ranking der Hochschulen (Abbildung 1) auf.

Abbildung 1: Adjustiertes Ranking von Universitäten im Studentenpisa-Test mit Studienfach „Wirtschaftswissenschaften“.



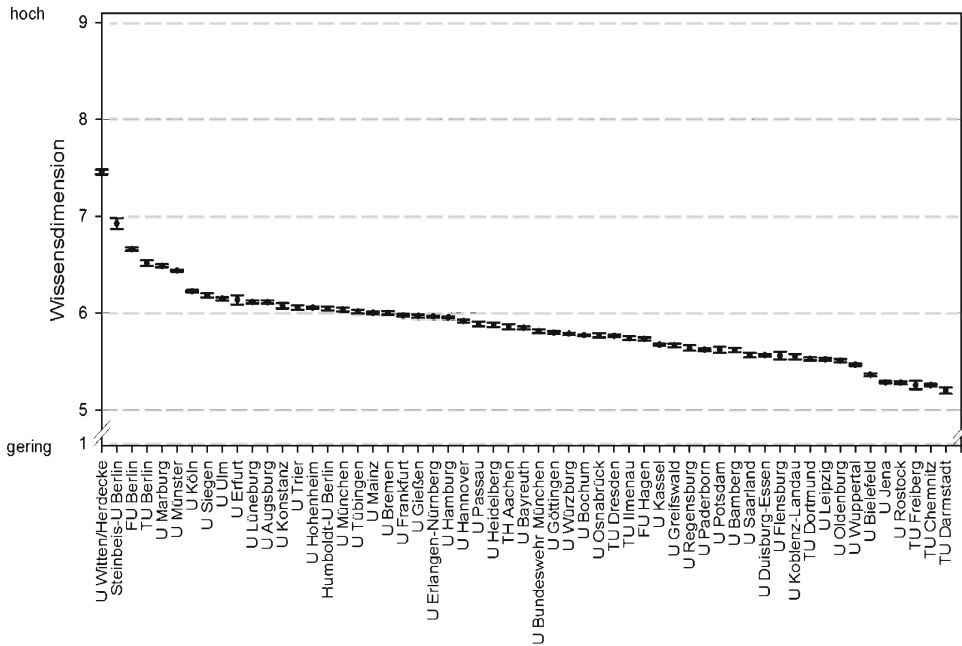
Mehrebenenanalyse der Antworten von Studierenden der Wirtschaftswissenschaften im Subtest „Allgemeinwissen in Wirtschaft“. Die Universitäten sind mit einem Vertrauensintervall von 95% (Fehlerbalken) von links (die im Wissenstest am besten abscheidenden Universitäten) nach rechts sortiert.

#### 4.5 Die Rangordnung

Im Weiteren wurden die Universitäten entsprechend ihrer mittleren Werte auf der Wissensdimension in eine Rangordnung gebracht (sortiert). Die vertikal dargestellten Fehlerbalken spiegeln die stichproben- und messfehlerbedingten Schwankungen des jeweiligen Werts einer Universität auf der Wissensdimension wider (95 Prozent Vertrauensintervall). Statistisch bedeutsam unterscheiden sich nur Universitäten im Allgemeinwissen in Wirtschaft, deren Fehlerbalken sich nicht überlappen (Goldstein & Healy, 1995). Wie oben angedeutet unterscheiden sich nahe beieinander liegende Hochschulen statistisch nicht signifikant. So gibt es beispielsweise zwar Rangunterschiede zwischen der FU Berlin (Rang 1) und der Universität Erlangen-Nürnberg (Rang 6). Die Konfidenzintervalle der beiden Universitäten überlappen sich jedoch, sodass die Rang- bzw. Mittelwertunterschiede keine realen Unterschiede, sondern nur Zufalls- oder Messfehlereffekte widerspiegeln. Rangordnungen von Hochschulen ohne Angabe eines Fehlerbalkens können demnach zu krassen Fehlschlüssen führen. Allenfalls lassen sich Gruppen von Universitäten finden, die sich statistisch signifikant unterscheiden. So unterscheidet sich eine ranghöhere Gruppe von Universitäten (FU Berlin, Universität Witten/Herdecke, Universität Erlangen-Nürnberg, Universität Hohenheim) von einer rang-

niederen Gruppe von Universitäten (Universität Wuppertal, Universität Oldenburg, Universität Jena, Universität Rostock).

Abbildung 2: Rohranking von Universitäten im Studentenpisa-Test mit Studienfach „Wirtschaftswissenschaften“.



Rangordnung der Rohmittelwerte bezüglich der Summenskala richtiger Antworten im Subtest „Allgemeinwissen in Wirtschaft“. Die Universitäten sind mit einem Vertrauensintervall von 95% (Fehlerbalken) von links (die im Wissenstest am besten abscheidenden Universitäten) nach rechts sortiert.

Abbildung 2 zeigt abschließend das Roh-Ranking auf der Grundlage der Summenskala des Wissenstests. Es wurden – wie dies üblicherweise getan würde – einfach die Mittelwerte je Universität über die Summenskala des Wissenstests berechnet. Diese ergibt sich als Summe der richtigen Antworten über die 9 Testitems. Ergänzend wurde das 95%-Konfidenzintervall (Fehlerbalken) auf der Grundlage des Standardfehlers des Mittelwerts bestimmt. Dieses Roh-Ranking weist zwar eine mäßige Korrelation mit dem adjustierten Hochschul-Ranking auf (Kendall's Tau-b von 0.59), aber die Fehlerbalken sind im Vergleich zu Abbildung 1 viel zu klein und suggerieren damit fälschlicherweise „statistisch signifikante“ Rangunterschiede zwischen einzelnen Universitäten.

## 5. Fazit

Insgesamt zeigen diese exemplarischen testtheoretischen Analysen, dass der Studententpisa-Test eine geeignete Grundlage für ein Hochschulranking darstellt. Jedoch sind infolge der hohen Wissensunterschiede innerhalb einer Hochschule allenfalls Rangunterschiede zwischen Gruppen von Universitäten interpretierbar, nicht Rangunterschiede zwischen einzelnen Universitäten. Für ein wissenschaftlich belastbares Hochschul-Ranking müssten weitere Korrekturen vorgenommen werden, beispielsweise durch den Einbezug weiterer Informationen über die Population, um durch eine post-stratifizierende Adjustierung Stichprobenverzerrungen noch besser korrigieren zu können oder durch den Einbezug von weiteren möglichen urteilsverzerrenden Faktoren (z. B. Abiturnote), um die Fairness des Hochschulvergleichs noch umfassender zu gewährleisten.

## Literatur

- Adams, R. J. & Wu, M. L. (2002). PISA 2000 Technical Report. OECD (www.pisa.oecd.org, 28.2.2010)
- Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. London: Guilford Press.
- Berghoff, S. & Federkeil, G. (2009). *The CHE Approach*. In C. Dehon, D. J. & C. Vermandele (Eds.), *Ranking Universities* (pp.41-63). Editions de L'Universite de Bruxelles.
- Biemer, P. P. & Christ, S. L. (2008). Weighting Survey Data. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 317-341). London: Lawrence Erlbaum.
- Cochran, W. G. (1977). *Sampling Techniques* (3<sup>rd</sup> ed.). New York: Wiley.
- Daniel, H.-D. & Hornbostel, S. (1993). Evaluation der Lehre: Sonderauswertung der SPIEGEL-Studie 1993 für Physik. *Physikalische Blätter*, 49(10), 903-906.
- Dill, D. & Soo, M. (2005). Academic Quality League Tables, and Public Policy: A Cross-national Analysis of University Rankings. *Higher Education*, 49, 495-533.
- Elton, L. (2004). Goodhart's Law and Performance Indicators in Higher Education. *Evaluation and Research in Education*, 18 (1-2), 120-128.
- Engel, U. & Krekler, G. (2001). Studienqualität – Über studentische Bewertungen und Rankings von Studienfächern einer Universität. In U. Engel (Hrsg.), *Hochschul-Ranking – Zur Qualitätsbewertung von Studium und Lehre* (S. 121-176). Frankfurt a. M.: Campus.
- Goldstein, H. & Healy, M. J. R. (1995). The Graphical Presentation of a Collection of Means. *Journal of the Royal Statistical Society, Series A*, 158(1), 175-177.
- Goldstein, H. & Leckie, G. (2008). School League Tables: What can they really tell us? *Significance*, 5(2), 67-69.
- Goldstein, H. & Spiegelhalter, D. (1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society, A*, 159, 385-433.
- Goodhart, C. (1981) Problems of Monetary Management: The UK Experience. In A. Courakis (Ed.) *Inflation, Depression and Economic Policy in the West* (p. 111-144). New Jersey: Mansell Publishing.
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43(4), 326-352.
- Holland, P. W., Dorans, N. J. & Petersen, N. S. (2007). Equating Test Scores. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics – Handbook of Statistics Vol 26* (pp. 169-203). London: Elsevier.
- Hornbostel, S. (2007). Theorie und Praxis von Hochschulrankings. In Statistisches Bundesamt (Hrsg.): *Statistik und Wissenschaft*, Band 11. Amtliche Hochschulstatistik und Hochschulrankings, 6-13.
- Hox, J. (2002). *Multilevel Analysis – Techniques and Applications*. London: Lawrence Erlbaum.
- Kroth, A. & Daniel, H.-D. (2008). Internationale Hochschulrankings – ein methodenkritischer Vergleich. *Zeitschrift für Erziehungswissenschaft*, 11, 542-558.

- Leberherz, C., Mohr, C., Henning, M. & Sedlmeier, P. (2005). Wie brauchbar sind Hochschulrankings? Eine empirische Analyse. *Zeitschrift für Pädagogik*, 51 (Beiheft), 188-208.
- Lohr, S. L. (2008). Coverage and Sampling. In E. D. de Leeuw, J. J. Hox, & D. A. Dillmann, (Eds.), *International Handbook of Survey Methodology* (pp. 97-112). London: Lawrence Erlbaum.
- Lubrano, M. (2009). A Statistical Approach to Rankings: Some Figures and Explanations for European Universities. In C. Dehon, D. Jacobs & C. Vermandele (Eds.), *Ranking Universities* (pp. 77-99). Editions de l'Université de Bruxelles.
- Lynn, P. (2008). The Problem of Nonresponse. In E. D. de Leeuw & J. J. Hox, & D. A. Dillmann, (Eds.), *International Handbook of Survey Methodology* (pp. 35-55). London: Lawrence Erlbaum.
- Marsh, H. (2007). Students' Evaluation of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In R. P. Perry & J. C. Smart (Eds.). *The Scholarship of Teaching and Learning in Higher Education: An Evidence-based Perspective* (pp. 319-384). New York: Springer.
- Mutz, R. & Daniel, H.-D. (2007). Entwicklung eines Hochschul-Rankingverfahrens mittels Mixed-Rasch-Modell und Mehrebenenanalyse am Beispiel der Psychologie. *Diagnostica*, 53(1), 3-17
- Rindermann, H. (2001). *Lehrevaluation. Einführung und Überblick zur Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen. Mit einem Beitrag zur Evaluation computerbasierter Unterrichts*. Landau: Empirische Pädagogik.
- Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*, 29(1), 185-203.
- Salmi, J. & Saroyan, A. (2007). League Tables as Policy Instruments: Uses and Misuses. *Higher Education Management and Policy*, 19(2), 1-38.
- Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis – An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
- Spiel, C. (2001). Der differentielle Einfluss von Bias-Variablen auf studentische Lehrveranstaltungsbewertungen. In U. Engel (Hrsg.), *Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre* (S. 61-82). Frankfurt a. M.: Campus.
- Statistisches Bundesamt (2009). *Studierende und Studienanfänger nach Land, Hochschule, Studienfach Wirtschaftswissenschaften und Geschlecht im WS 2008/2009*. Tabelle VI B – HS.
- Usher, A. & Savino, M. (2007). A Global Survey of University Rankings and League Tables. *Higher Education in Europe*, 32(1), 5-15.
- Van der Wende, M. (2008). Rankings and Classifications in Higher Education: A European Perspective. In J. C. Smart (Eds.), *Higher Education: Handbook of Theory and Research* (pp. 49-71). New York: Springer.
- Wissenschaftsrat (2008). *Empfehlungen zum Forschungsrating. Wettbewerb im deutschen Hochschulsystem*. Mai 2008 (Drs. 8485-08). Köln.