JOURNAL OF THE PROGRAMME
ON INSTITUTIONAL MANAGEMENT IN HIGHER EDUCATION

Higher Education Management and Policy

Volume 19, No. 2

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

ISSN 1682-3451
Higher Education Management and Policy
Volume 19, No. 2
© OECD 2007

Universities on the Catwalk: Models for Performance Ranking in Australia

by

Hamish Coates
Australian Council for Educational Research (ACER), Australia

National and international rankings of institutional performance are playing a growing role in contemporary higher education. It is critical that researchers develop pragmatic, educationally sensitive and methodologically informed approaches for managing this aspect of higher education. This paper compares three approaches for modelling key indicators which underpin a national evaluation of university education in Australia: rankings of aggregate institutional performance; comparisons of institutional change over time; and performance variations within fields of education. The results show that simple institution-level aggregations are misleading, and that contemporary analytical methods must be used to account for the influence of fields of education. More broadly, the findings expose the need for a more robust methodological development of university rankings.

Investigating the modelling of university performance

Despite its critics and inherent difficulties, it seems very likely that university rankings are here to stay. Higher education markets are becoming more open and competitive, with increasing calls for information about quality and effectiveness. Government, business, potential students, the general public and institutions themselves want more and better information to help differentiate varying levels of quality and performance. It is critical, as such, that researchers develop pragmatic, educationally sensitive and methodologically informed approaches for managing this increasingly prominent aspect of higher education.

Much energy has been invested in producing rankings over the last decade. National rankings have been produced to capture research and educational performance (US News, 2006; Hobsons, 2006; Siwinski, 2002; Williams and Van Dyke, 2005; DEST, 2005). Two prominent international rankings (THES, 2004; IHE, 2004) have emphasised university research, although the OECD has begun conversations about possible cross-national assessments of student performance (Ischinger, 2006). In Canada, an innovative attempt is underway to develop a "composite learning index" that represents the current state of learning across the country (Cartwright *et al.*, 2006). As Guthrie (1993) portended, education systems may well be on the road to a "Dow Jones" index.

While not designed explicitly as a ranking mechanism, attention in Australia has been focused on development and administration of the Learning and Teaching Performance Fund (LTPF) (Nelson, 2003). In 2003, the Australian government signalled its interest in evaluating and rewarding higher education teaching and learning at the national level. The LTPF was developed from 2003 to 2005 to "reward those institutions which best demonstrate excellence in teaching and learning" (Nelson, 2003, p. 29). Three annual funding rounds will have been conducted by the end of 2007, distributing around AUD 250 million to a selection of 38 eligible Australian institutions. The results have also been used to generate "learning and teaching" rankings of institutions. The LTPF is an interesting policy initiative, not least because it includes and affects an entire national system.

Such ranking activities generate substantial discussion and debate. As part of this, ranking methodology is emerging as a significant area of higher education research. One area of focus has been the policy contexts which surround ranking (Merisotis, 2002a; Merisotis and Sadlak, 2005; Yonezawa *et al*., 2002; Cai Liu and Cheng, 2005). Work has also focused on understanding the nature and selection of indicators, which is important given that rankings are ultimately only as valid as the data on which they are based (Coates, 2006; Van Dyke, 2005; Clarke, 2002). Developing appropriate statistical approaches for modelling indicator data is a further growing concern of ranking methodology research (Clarke, 2002; Van Dyke, 2005; Filinov and Ruchkina, 2002). Researchers have also considered standards for how reports and data are best used (Clarke, 2005; IREG, 2006), and early meta-analytic work has been done on the development of rankings frameworks and typologies (Usher and Savino, 2006; Dill and Soo, 2003; Merisotis, 2002b; OECD, 2006). Relatedly, international work has been initiated by the Institute for Higher Education Policy and UNESCO-CEPES (the European Centre for Higher Education) (Merisotis and Sadlak, 2005; Merisotis, 2002b) to develop an International Rankings Expert Group to monitor ranking activities.

This paper contributes to the methodological discussion of university rankings by investigating alternative approaches for modelling key indicators which underpin a national evaluation of university education. It develops findings based on the analysis of data used in the Australian Learning and Teaching Performance Fund. The findings are used to explore approaches to the large-scale evaluation of university education which are of relevance to institutional researchers around the world. Large-scale analysis of educational performance is invariably high-stakes, and conducting data analyses in valid and appropriate ways is critical.

The teaching and learning focus of this paper is largely incidental to its main methodological intent, but does add an extra dimension to the analysis. As suggested in the above overview, rankings have tended to focus on the research rather the educational function of universities. This raises interesting questions about the “research/teaching nexus” and determinants of higher education quality, which lie beyond the scope of this paper. Indirectly, however, this paper does offer a timely juxtaposition to the general focus on research in discussions of higher education rankings.

The broader intention of this paper is to stimulate awareness of state-of-the-art statistical techniques in higher education policy circles. While aspects of large-scale evaluation are relatively new to higher education, methods used to monitor educational effectiveness have been widely used and rigorously tested over many years. Examples include school effectiveness research (Woodhouse and Goldstein, 1988; Bottani and Tuijnman, 1994; Hill and Rowe, 1998) and large-scale studies of educational achievement (OECD, 2005; NCES, 2006). In contrast, only a few methodological studies have been published in higher education journals (Guarino et *al.,* 2005; Rocki, 2005). While a certain re-learning is required when techniques are transported into new contexts with new audiences, the use of contemporary analytical methods to maximise the validity of large-scale evaluations of university education is imperative.

The data and its context

The paper analyses data from the Course Experience Questionnaire (CEQ). The CEQ is conducted in a census of all coursework graduates administered around four months after graduation. The data is collected by participating institutions, then compiled and analysed by national agencies. The national reports are distributed to institutions, the public and government, and provide baseline figures for many intra- and cross-institutional activities. The results in this paper are based mainly on data collected from the 2005 census of 191 998 coursework graduates at 42 institutions which returned 98 138 usable responses (GCA/ACER, 2006).

The CEQ was developed in the early 1990s (Ramsden, 1991) and a series of new scales were added a decade later (McInnis *et al.*, 2001). While the full CEQ measures 11 qualities of the educational experience, this paper focuses on two scales and a single item indicator that have been administered by all Australian universities since 1992: Good Teaching Scale (GTS), Generic Skills Scale (GSS) and Overall Satisfaction Item (OSI). Since 2005, these three indicators have been included in the seven analysed for the LTPF.

For current purposes, the 13 items spread across the GTS, GSS and OSI have been combined to form a single Quality of Teaching and Skills (QTS) scale. The QTS has good measurement properties. Its alpha reliability is 0.91, and its congeneric reliability estimate (Werts *et al.*, 1978; Reuterberg and Gustafsson, 1992) is 0.95. Congeneric measurement modelling affirmed the construct validity of the QTS scale. All estimated parameters are statistically significant and have a mean standardised value of 0.78. The root mean square error of approximation for the model is 0.07, the non-normed fit index is 0.89 and the goodness of fit index is 0.98. In summary, the QTS provides a consistent composite measure of selected aspects of university education.

All Quality of Teaching and Skills items are answered using a five-point response scale which runs from “strongly disagree” to “strongly agree”. The analyses reported in this paper are based on a rescoring of the five response categories to –100, –50, 0, 50 and 100. This metric expands the range of the reporting scale and eliminates the need to analyse decimal-place differences. For parsimony, the QTS scores analysed in this paper have been calculated using simple summative methods even though other psychometric methods would produce more reliable measures.

QTS scores can be interpreted in a range of ways. While the data is collected as part of a census, the extent of survey non-response makes it appropriate to use statistical methods to analyse the sample of secured data.

Given the large number of QTS responses, the standard errors at the respondent level are around 0.5 and hence a difference of just over 1.5 on the –100 to +100 scale is likely to be statistically significant. This statistical significance is an artefact of the large number of observations, however, and small differences are likely to carry little meaning in practice. In large-scale surveys, statistical significance is an artefact of the large number of observations and many statistically significant differences are likely to carry little meaning in practice. A preferable approach is based on "effect size" (Cohen, 1969). Measures of effect size indicate the magnitude of the difference between two scores in standard deviation units which are independent to the number of observations. By convention, a difference of 0.2 standard deviation units is considered to be a small effect, a difference of 0.5 units a medium effect and 0.8 units a large effect size. As the standard deviation of the QTS scores is 34.5, differences of 12 points or more may be of interest as they represent a margin of at least a fifth of a standard deviation.

Three approaches for identifying educational quality

This paper explores three approaches for using data from the Course Experience Questionnaire to identify fields of educational performance in Australian higher education. It analyses aggregate institutional performance, change in institutional performance over time and institutional performance within broad fields of education. Results from these analyses are considered in terms of their implications for ranking institutional performance.

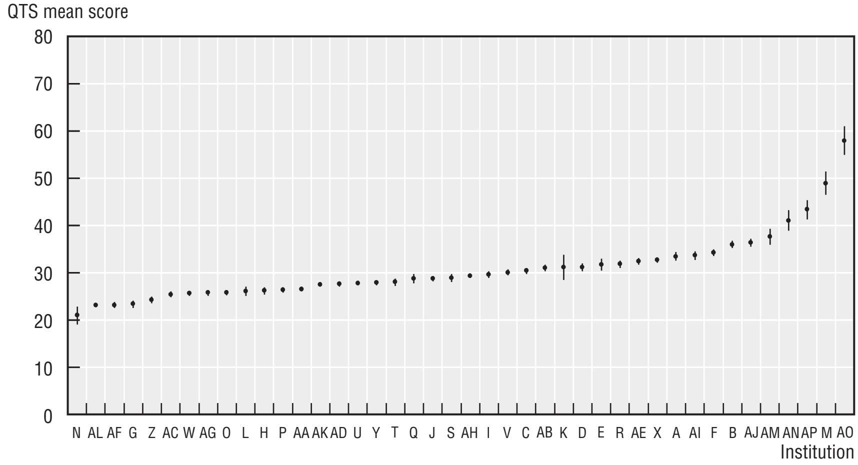
Aggregate institutional performance

The first approach tests the value of aggregating information from indicators such as the CEQ to the institutional level. There is often a strong interest in reviewing CEQ results at the institutional level. Institutions are often viewed and branded as relatively homogeneous corporate entities, even though they may in fact comprise heterogeneous and dynamic educational communities.

Figure 1 presents each institution's aggregate QTS score sorted in ascending order with 95% confidence bands. Each band marks out the interval which is very likely to include the true population mean score for each institution. The confidence bands have been adjusted for pairwise comparisons and the finite nature of the population. The 42 institutions have been coded randomly from A to AP for reporting purposes.

Figure 1 shows a large number of differences between institutions from a purely statistical perspective. A few "stand out" institutions have high QTS scores, around half have "above average" scores (greater than 28.3), while there is less differentiation between institutions towards the lower end of the

Figure 1. **Quality of Teaching and Skills scores by institution, 2005**



scale. These differences are marginal when considered in terms of the rescaled response category units of 50, however such variation is often treated seriously by stakeholders and funding agencies.

Fewer differences between institutions exist in terms of meaningful effect size. Only a few have score differences of 12 points or more, such as institutions A and AN. From an effect size perspective, the results in Figure 1 expose only around two to three different levels of institutional performance. While various groupings are possible, institutions M and AO could be placed in an upper band, A to AP in a middle band, and N to X in a lower band. With further psychometric modelling, such levels could be differentiated into qualitatively interpretable performance thresholds and hence quality benchmarks. While beyond the scope of this paper, such modelling would offer a much more sophisticated alternative to the use of statistical methods alone.

Institutional change over time

Measures of "improvement" or "value added" are the most powerful indicators of educational performance. Determining improvement, however, requires identifying a baseline against which it can be assessed. Since individuals complete the CEQ only once after their course, the cross-sectional nature of the data makes estimating net effects for individual students or graduates impossible. However, calculating improvement at the institutional or field of education level in terms of change over years is possible. It must be stressed that measures of "improvement" and "value added" are measures of growth and performance, and are not the same as "industry" or "effort".

Linear regression was used to estimate institutional change relative to performance set by population expectations. Such estimates are called "residual change scores". Residual change scores have a number of desirable properties and are used widely to measure educational effectiveness (Glass and Hopkins, 1996; Goldstein, 1995; Woodhouse and Goldstein, 1988). They are preferable to change or difference scores calculated using simple subtractive methods, as the reliability of these simpler measures tends to be low and they are perturbed by floor and ceiling effects (Linn, 1988; Cronbach and Furby, 1970).

Regression analysis exposed a strong linear relationship between the 2004 and 2005 scores. The 2004 scores account for 82.2% of the variation in the 2005 scores, and the correlation between the scores is 0.91. The relationship between the scores and the linear line of best fit is shown in Figure 2. The line of best fit represents expected performance. Institutions represented by points above the line have performed above expectation, and those represented by points below the line have performed below expectation.

Figure 2. **Quality of Teaching and Skills mean scores by institution, 2004-05**

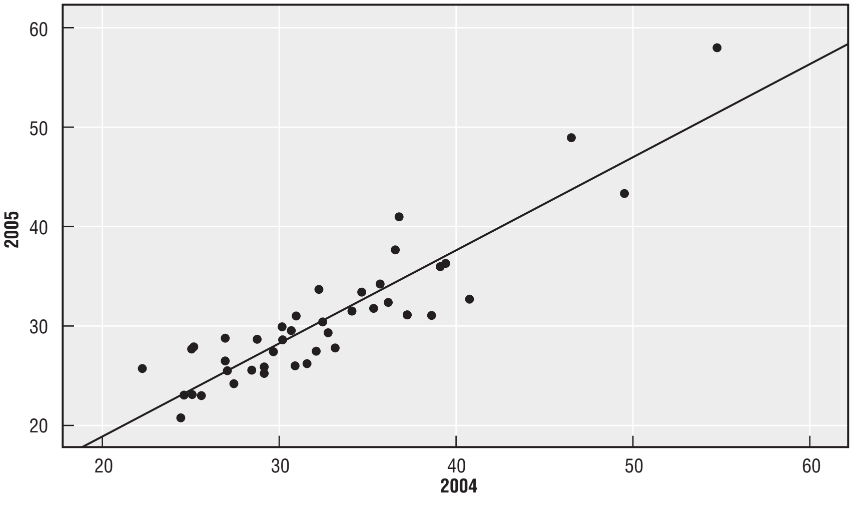
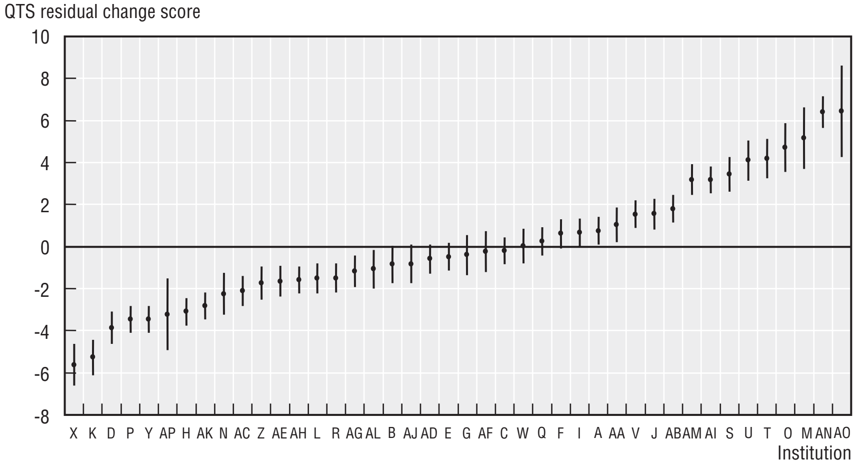


Figure 3 shows the residual change scores for each institution along with confidence bands sorted from lowest to highest. There appear to be three different types of change. Figure 3 exposes about 15 institutions with greater than average change, around 10 with score changes in line with expectations, and about 17 with 2005 QTS scores less than expected given their 2004 performance. Differentiation is greater among institutions at the lower and upper ends than in the middle of the distribution.

Figure 3 also enables analysis of the rate of QTS score change across years. While institutions V to AO have a relatively fast rate of change, Z to AA are changing little, while X to AC are sliding backwards. Institutional performance could be measured in terms of such change gradients, which indicate the extent of improvement or decline in teaching quality as measured by the QTS. Absolute performance aside, there is often much value in educational organisations which are experiencing conditions of growth and productivity.

Figure 3. **Quality of Teaching and Skills residual change scores by institution, 2004-05**



It is interesting to note the differences between the measures of aggregate institutional performance and the residual change scores. The overall correlation between the rankings is quite low at only 0.26. Despite this, there are telling patterns in the lists. While institution AO is at the top of both rankings, only three common institutions rank in the top five of each. Only four common institutions rank in the top ten. While these rankings differ, they do indicate that it is possible to have high levels of both aggregate performance and change. They also indicate that, independent of important questions about the appropriateness of institutional aggregations, the measurement of absolute performance and of change in performance across consequence years are two complementary approaches for reviewing educational quality using CEQ data.

Performance within fields of education

In reviews of educational performance, employing analytical approaches which are sensitive to the phenomena being analysed is critical. While higher education institutions tend to have complex and idiosyncratic structures, CEQ scores in large part reflect the perceptions of graduates who learned within fields of education, within institutions. To minimise bias, therefore, it is essential that analyses account for the hierarchical structure inherent in the data.

Applying a single-level analytical perspective, such as in the analysis of aggregate institution scores, can cause a range of problems. First, single-level analyses ignore the effects of clustering and treat all observations as independent. In doing this, they overestimate the number of unique observations and hence the amount of information being analysed. Overestimating the information being analysed leads to underestimating standard errors, which in turn makes the identification of spurious differences more likely.

Second, and most importantly, it is fallacious to assume that relationships identified at aggregate levels hold for subgroups or for individual members. In technical terms, single-level analyses run the risk of committing the "ecological fallacy" (Robinson, 1950). While a whole may be the sum of its parts, it may not be equivalent to them. Ignoring the heterogeneity among elements within groups not only leads to the misapplication of general structure over group particularity but also ignores, as suggested above, the rich possibilities made available by studying such difference. Put simply, aggregate pictures of educational performance misrepresent the diversity within institutions.

A further reason for taking a multilevel perspective is that higher education institutions vary in their composition. Not all universities, for instance, include medical, physiotherapy or education schools. Drawing comparisons between organisations with different structures and academic units can be misleading, perhaps even more so than representing an institution by a single number. Disaggregated reports require the analysis of more information, but they enable interpretations to be drawn between comparable faculties and schools.

Covariance analyses of Course Experience Questionnaire data (GCA/ACER, 2006) indicate that both the institution itself and the field of education underpin patterns of score variation. Analyses (specifically multilevel variance components modelling) of graduates within fields of education within institutions suggest that individual respondents account for 94.7% of variation in QTS scores, the broad field of education accounts for 3.1%, while institutions account for only 2.2%. The results of this variance decomposition are pivotal. They indicate that the field of education causes more variation in educational performance, as measured by the QTS, than does an institution. This suggests that analyses of educational performance based on CEQ data should not ignore the field of education. While parsimony is important, explanatory statistical models should aim to explain as much meaningful variation in the data as possible.

This important point is exemplified in Figure 4, which shows QTS mean scores and 95% confidence bands for three institutions across the ten main broad fields of education. The selected institutions are the same type of university, share many common characteristics and have a large number of CEQ responses in each field of education. They are also spread across the distributions of institutions shown in Figures 1 and 3. What Figure 4 shows is that the institutions perform in different ways depending on the field of education. Institution AK, which has QTS scores well below those of institution X in the Information Technology field, has a much higher QTS score in the Agriculture, Environmental and Related Studies field. Institution X appears to have consistently high scores, although the difference is marginal in six of the ten fields.

Figure 4. **Quality of Teaching and Skills mean scores for sample institutions by field of education, 2005**

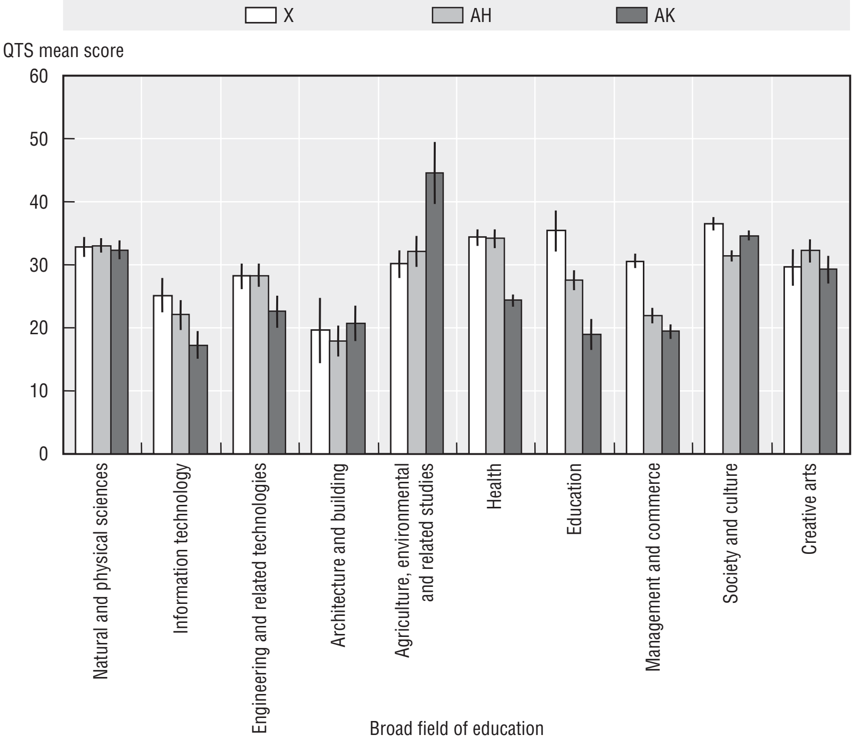
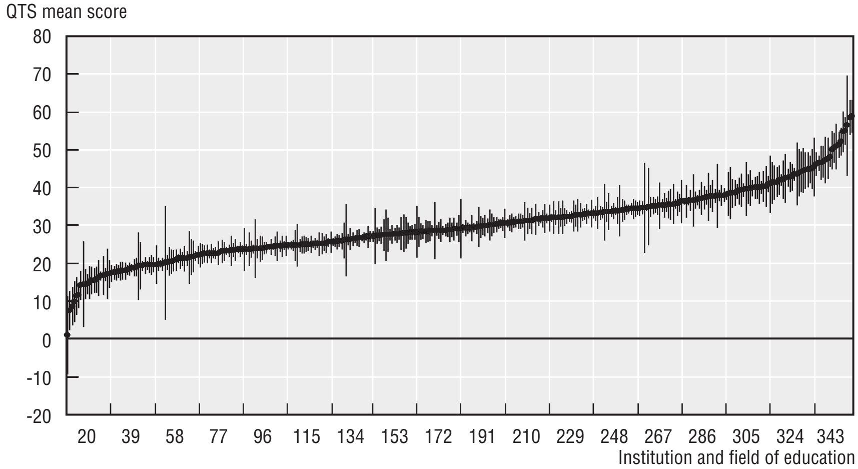


Figure 5 presents a national picture of the performance of three institutions within fields of education. It shows QTS mean scores with greater than ten responses for each combination of institution and broad field of education. The labels on the horizontal axis reflect the rank order of the observation. The mean scores have been sorted in ascending order and are shown with confidence

Figure 5. **Quality of Teaching and Skills mean scores by institutions and fields of education, 2005**



bands adjusted for the finite population and pairwise comparisons. The plot exposes a wide range in performance across fields of education in Australian institutions as measured by the QTS, in terms of both statistical significance and effect size.

A summary of the different rankings

In Table 1, the 42 institutions are sorted in order of descending QTS mean score or residual change score. This presentation does not include information about sampling error or effect size, therefore many of the differences between institutional QTS scores may be inconsequential.

Despite its limitations, Table 1 provides a useful summary of the analyses explored in this paper. It shows, for instance, a considerable variation in institutional order across the lists, a result which reinforces the need to use robust analytical methods. There is also variation in the amount of detail given by each approach. While the aggregate institutional and annual change approaches provide a result for each institution, the multilevel approach provides a result for each field of education taught at each institution. This latter approach offers more sensitive and accurate information to assist subsequent interpretations of educational performance.

Developing ranking methodology for higher education

This paper has explored different ways in which indicator scores might be used to measure variations in the quality of university education. By investigating different approaches, the analysis has sought to advance

Table 1. **Institutional performance rankings, 2005**

		Broad field of education									
Whole institution	Annual change	Natural and physical sciences	Information technology	Engineering and related technologies	Architecture and building	Agriculture, environmental and related studies	Health	Education	Management and commerce	Society and culture	Creative arts
AO	AO	AE	AM	M	AC	AJ	AM	AO	M	AO	AJ
M	AN	AJ	A	B	J	D	AN	J	AP	M	AP
AP	M	N	B	N	AF	AE	M	M	K	AM	B
AN	O	S	AP	E	C	S	C	AJ	B	AN	D
AM	T	A	AE	F	R	AK	F	AN	AN	F	R
AJ	U	I	S	R	AD	AC	T	U	AM	AP	Q
B	S	AM	AB	AE	F	AG	E	R	F	G	Y
F	AI	H	F	A	AA	A	X	AI	AI	B	A
AI	AM	R	P	AC	U	V	AH	X	I	A	F
A	AB	D	AI	AB	O	E	H	D	X	R	H
X	J	L	J	AH	AJ	AI	B	W	Q	AB	E
AE	V	F	R	X	V	I	L	E	R	D	AB
R	AA	B	X	T	W	T	AJ	C	A	V	P
E	A	AC	I	I	AK	U	V	Z	AJ	AJ	C
D	I	E	E	J	X	J	P	AA	AB	Q	AE
K	F	V	D	U	S	AF	A	V	AD	X	AH
AB	Q	C	G	O	AH	C	AI	H	V	Y	L
C	W	U	AH	C	AL	AM	J	AD	J	E	O
V	C	G	AJ	G	L	AH	AD	AH	U	C	AF
I	AF	T	AD	D	T	AA	AG	G	S	I	W
AH	G	AG	C	W		P	I	Q	Y	AI	X
S	E	AH	V	AK		AL	D	A	O	H	AM
J	AD	X	AA	V		X	W	AF	AA	AK	AA
Q	AJ	AA	W	H		B	AA	I	AE	AF	AK
T	B	AK	AC	AD		Y	AB	P	E	J	N
Y	AL	AF	U	K		Z	R	L	AC	S	AI
U	AG	W	O	AL		O	AC	AC	D	O	AL
AD	R	AL	AF	AA		K	Y	AB	AG	L	J
AK	L	AB	AK	Y		AD	AK	AG	C	AE	V
AA	AH	J	AG	AG		AB	S	Y	AL	AD	U
P	AE	Z	AL	AF			Q	S	L	U	AD
H	Z	AD	H	S			G	AP	P	AH	I
L	AC	Y	Y				O	O	W	T	S
O	N	O	Q				AL	AK	AH	Z	AG
AG	AK	P	L				U	AL	T	N	AC
W	H		Z				AF	N	G	AG	Z
AC	AP		T				Z	B	Z	AA	G
Z	Y		N				N	T	AK	W	T

Table 1. **Institutional performance rankings, 2005** *(cont.)*

		Broad field of education									
Whole institution	Annual change	Natural and physical sciences	Information technology	Engineering and related technologies	Architecture and building	Agriculture, environmental and related studies	Health	Education	Management and commerce	Society and culture	Creative arts
G	P								H	AL	
AF	D								AF	AC	
AL	K								N	P	
N	X										

understanding of the methodology which underpins ever increasing large scale evaluations of educational quality. The analysis has shown that, as expected, different approaches to analysing indicator data produce different results. This observation is simple but not trivial, for the consequences of such quality determinations can often be enormous. As recent national and international exercises demonstrate, rankings can have significant effects on higher education funding, perceptions of quality, enrolments and trade.

The modelling in this paper has shown that it is essential to use forms of analysis which provide valid, reliable, efficient and informative results. Simple institution-level indicator scores alone are unlikely to achieve this goal. While they provide insight into overall institutional performance and allow an efficient means of reviewing change over time, they conflate the important patterns of variation which are due to the field of education. A multilevel form of analysis which reflects the reality that students learn within fields of education within institutions enables the production of more robust performance estimates. The estimates also provide more valuable information to institutions, managers students and the public, as they offer evidence at the level at which educational decisions are often made.

Ranking methodology is a relatively new field of inquiry in higher education. Much ongoing work is needed to explore other issues central to producing university rankings. While this paper has focused on the modelling of indicator data, research on rankings should be multifaceted and consider a range of practical, methodological and substantive issues.

Further education-focused and policy-level reviews should be conducted to examine which indicators are best used in large-scale university classifications and rankings. This may require the development of data on learning processes and outcomes to augment the systemic collections established over the last few decades. Rankings of university quality must not rely on indicators which are simply ready-to-hand, but rather reflect the important aspects of university education.

Researchers need to consider what psychometric methods should be used to produce indicator scores. While rankings are often composed by analysts with expertise in secondary data analysis, it is critical that measurement considerations are not overlooked. Work is required to identify scaling procedures which extract the maximum amount of meaningful variation in indicator scores. Funding and educational decisions which flow from institutional rankings may be based on little more than chance if variations in indicator scores are unreliable and reflect random measurement error.

An important progression will involve linking ranking processes with more general higher education research. While the US National Survey of Student Engagement (NSSE, 2006) exemplifies the projection of research into practice, most rankings connect in very tangential ways if any with what is known about the nature and development of quality in university education. While reputation and resource indices frequently factor into rankings, for instance, empirical research (Pascarella and Terenzini, 2006) has shown that the relationship between these and the effectiveness of undergraduate education is low. At the same time, no rankings include the results of psychometrically validated student assessments of subject-specific knowledge or generic skills.

The development of frameworks and typologies will play an important part in enhancing research on rankings. Frameworks might provide classification of the different types of rankings, of the composition of different rankings, of the relevance of rankings for different institutions, or of how analysts and consumers might equate different rankings. They can provide a lens for reviewing the contexts, nature and implications of different rankings, and a structure against which progress in developing rankings can be planned and measured.

Rankings are attractive because they provide easily consumed information on selected aspects of higher education quality. Such simplicity can be problematic, however, as rankings which are computed or used in inappropriate ways can cause much harm to institutions and national systems. While numbers themselves are often context-free, it is critical that they are used in contextually sensitive ways. One of the most important areas for development will be to define standards for the appropriate reporting and use of rankings. Given the national and international scope of most rankings, such standards are likely to grow through ongoing discussion and debate.

Universities at the top of rankings can leverage much more than they should from the small differences which often place them there. University rankings are persuasive and increasingly form part of everyday conversations about higher education. Therefore, and as suggested in this paper, researchers must develop robust methodologies to assure the validity of such lists.

Acknowledgements

I am grateful to Graduate Careers Australia (GCA) for permission to include analyses based on the national data in this paper. Professor Ross Williams and Dr. Gary Marks provided helpful feedback on earlier drafts of this paper.

The author:

Dr. Hamish Coates
Australian Council for Educational Research (ACER)
347 Camberwell Road
Camberwell, Victoria, 3124
Australia
E-mail: *coatesh@acer.edu.au*

References

Bottani, N. and A. Tuijnman (1994), “International Education Indicators: Framework, Development and Interpretation”, in A. Tuijnman and N. Bottani (eds.), *Making Education Count: Developing and Using International Indicators,* OECD, Paris.

Cai Liu, N. and Y. Cheng (2005), “The Academic Ranking of World Universities”, *Higher Education in Europe*, Vol. 30, No. 2, pp. 127-136.

Cartwright, F., J. Mussio and C. Boughton (2006), *Developing a Composite Learning Index: A Framework,* Canadian Council on Learning, Ottawa.

Clarke, M. (2002), “Some Guidelines for Academic Quality Rankings”, *Higher Education in Europe,* Vol. 27, No. 4, pp. 443-459.

Clarke, M. (2005), “Quality Assessment Lessons from Australia and New Zealand”, *Higher Education in Europe,* Vol. 30, No. 2, pp. 183-197.

Coates, H. (2006), “Excellent Measures Precede Measures of Excellence”, paper presented at the Australian Universities Quality Forum, 6 July, Perth.

Cohen, J. (1969), *Statistical Power Analysis for the Behavioral Sciences,* Academic Press, New York.

Cronbach, L.J. and L. Furby (1970), “How Should We Measure ‘Change’: Or Should We?”, *Psychological Bulletin,* Vol. 74, pp. 68-80.

DEST (Department of Education, Science and Training) (2005), *Learning and Teaching Performance Fund,* DEST, Canberra.

Dill, D. and M. Soo (2003), “Is There a Global Definition of Academic Quality? A Cross-national Analysis of University Ranking Systems”, paper presented at the conference of International Network of Quality Assurance Agencies in Higher Education (INQAAHE), 17 April, Dublin.

Filinov, N.B. and S. Ruchkina (2002), “The Ranking of Higher Education Institutions in Russia: Some Methodological Problems”, *Higher Education in Europe,* Vol. 27, No. 4, pp. 407-421.

GCA/ACER (Graduate Careers Australia/Australian Council for Educational Research) (2006), *Graduate Course Experience, 2005: The Report of the Course Experience Questionnaire (CEQ),* GCA, Parkville.

Glass, G.V. and K.D. Hopkins (1996), *Statistical Methods in Education and Psychology,* Allyn and Bacon, Boston.

Goldstein, H. (1995), *Multilevel Statistical Models*, Edward Arnold, London.

Guarino, C., *et al.* (2005), “Latent Variable Analysis: A New Approach to University Ranking”, *Higher Education in Europe,* Vol. 30, No. 2, pp. 147-165.

Guthrie, J. (1993), “Do America’s Schools Need a Dow Jones Index?”, *Phi Delta Kappa*, Vol. 74, pp. 523-528.

Hill, P.W. and K.J. Rowe (1998), “Modeling Student Progress in Studies of Educational Effectiveness”, *School Effectiveness and School Improvement*, Vol. 9, No. 3, pp. 310-333.

Hobsons (2006), *The Good Universities Guide 2007: Universities and Private Colleges*, Hobsons Australia, Melbourne.

IHE (Institute of Higher Education) (2004), *Academic Ranking of World Universities – 2004,* IHE, Shanghai Jiao Tong University, Shanghai.

IREG (International Rankings Expert Group) (2006), *Berlin Principles on Ranking of Higher Education Institutions*, CHE (Centre for Higher Education Development), Gütersloh.

Ischinger, B. (2006), “Higher Education for a Changing World”, *OECD Observer*, No. 255, May.

Linn, R.L. (1988), “Change Assessment”, in J.P. Keeves (ed.), *Educational Research Methodology and Measurement: An International Handbook*, Pergamon, Oxford.

McInnis, C., *et al.* (2001), *Development of the Course Experience Questionnaire*, Department of Employment, Training and Youth Affairs, Canberra.

Merisotis, J. (2002a), “On the Ranking of Higher Education Institutions”, *Higher Education in Europe,* Vol. 27, No. 4, pp. 361-363.

Merisotis, J. (2002b), “Summary Report of the Invitational Roundtable on Statistical Indicators for the Quality Assessment of Higher/Tertiary Education Institutions: Ranking and League Table Methodologies”, *Higher Education in Europe,* Vol. 27*,* No. 4, pp. 475-480.

Merisotis, J. and J. Sadlak (2005), “Higher Education Rankings: Evolution, Acceptance and Dialogue”, *Higher Education in Europe,* Vol. 30, No. 2, pp. 97-101.

NCES (National Center for Education Statistics) (2006), *The NAEP 1998 Technical Report,* NCES, Washington, DC.

Nelson, B. (2003), *Our Universities: Backing Australia’s Future,* Department of Education, Science and Training, Canberra.

NSSE (National Survey of Student Engagement) (2006), “National Survey of Student Engagement”, NSSE, Bloomington.

OECD (2005), *PISA 2003 Technical Report,* OECD, Paris.

OECD (2006), “Institutional Diversity: Rankings and Typologies in Higher Education”, international workshop organised by IMHE and *Hochschulrektorenkonferenz*, 4-5 December, Bonn, *www.oecd.org/document/54/0,2340,en_2649_35961291_38046966_1_1_1_1,00.html*.

Pascarella, E.T. and P.T. Terenzini (2005), *How College Affects Students: A Third Decade of Research,* Jossey-Bass, San Francisco.

Ramsden, P. (1991), "A Performance Indicator of Teaching Quality in Higher Education: The Course Experience Questionnaire", *Studies in Higher Education*, Vol. 16, No. 2, pp. 129-150.

Reuterberg, S. and J. Gustafsson (1992), "Confirmatory Factor Analysis and Reliability: Testing Measurement Model Assumption", *Educational and Psychological Measurement,* Vol. 52, pp. 795-811.

Robinson, W.S. (1950), "Ecological Correlations and the Behavior of Individuals", *American Sociological Review,* Vol. 15, pp. 351-357.

Rocki, M. (2005), "Statistical and Mathematical Aspects of Ranking: Lessons from Poland", *Higher Education in Europe,* Vol. 30, No. 2, pp. 173-181.

Siwinski, W. (2002), "Perspektywy – Ten Years of Rankings", *Higher Education in Europe*, Vol. 27, No. 4, pp. 399-406.

THES (Times Higher Education Supplement) (2004), *World University Rankings,* THES, London.

US News (US News and World Report) (2006), *America's Best Colleges 2007,* US News and World Report Inc., Washington, DC.

Usher, A. and M. Savino (2006), *A World of Difference: A Global Survey of University League Tables,* Educational Policy Institute, Toronto.

Van Dyke, N. (2005), "Twenty Years of University Report Cards", *Higher Education in Europe,* Vol. 30, No. 2, pp. 103-125.

Werts, C.E., *et al.* (1978), "A General Method of Estimating the Reliability of a Composite", *Educational and Psychological Measurement,* Vol. 38, pp. 33-38.

Williams, R. and N. Van Dyke (2005), *Melbourne Institute Index of the International Standard of Australian Universities, 2005,* University of Melbourne, Melbourne Institute of Applied Economic and Social Research, Melbourne.

Woodhouse, G. and H. Goldstein (1988), "Educational Performance Indicators and LEA League Tables", *Oxford Review of Education*, Vol. 14, No. 3, pp. 301-320.

Yonezawa, A., I. Nakatsui and T. Kobayashi (2002), "University Rankings in Japan", *Higher Education in Europe*, Vol. 27, No. 4, pp. 373-382.