Are National-Level Research Evaluation Models Valid, Credible, Useful, Cost-Effective, and Ethical?

Chris L. S. Coryn and Michael Scriven
The Evaluation Center, Western Michigan University
School of Behavioral and Organizational Sciences, Claremont Graduate University

The evaluation of government-financed research has become increasingly important in the last few decades in terms of increasing the quality of, and payoff from, the research that is done, reducing the cost of doing it, and lending public credibility to the manner in which research is funded. But there are very large differences throughout the world in the extent to which systems used promote these results. This paper briefly presents the dimensional results of a study designed to comparatively evaluate the national-level research evaluation models in sixteen countries on five merit-defining dimensions.

In the last few decades the evaluation of research has become a high-stakes enterprise. With increasing political governance and federal budgets often in the billions, the livelihood of individual researchers, research groups, departments, programs, and entire institutions often swing in the balance. Simultaneously, it has been recognized that many of the longstanding principles and practices often lead to poor decisions about the actual or prospective merits of researchers and their research (Coryn, 2006, 2007).

This paper expands upon our work which examines the quality of national-level research evaluation models. Some of the details which exceed the scope of our recent paper due to appear in the *American Journal of Evaluation* (Coryn, Hattie, Scriven, & Hartmann, 2007) are presented here. This research comparatively evaluated the national models used to evaluate research and allocate research funding in 16 countries. Each of the models was rated on more than 25 quality indicators by two independent, blinded panels of professional researchers and evaluators in two countries (the United States and New Zealand). The indicators were used as observed, measurable aspects of five latent merit-defining criteria intended to represent a high-quality national research evaluation system. These dimensions were: validity, credibility, utility, cost-effectiveness, and ethicality.

The Countries

As shown in Figure 1, the 16 nations included in the study were: Australia; Belgium; the Czech Republic; Finland; France; Germany; Hong Kong; Hungary; Ireland; Japan; the Netherlands; New Zealand; Poland; Sweden; the United Kingdom; and the United States.

ISSN 1556-8180
October 2007

Validity

A good research evaluation model should produce conclusions that are logically correct and justifiable; that is, a good model should be valid. The average difference between the two panels' weighted validity scores was 0.81%; $t(15) = .34$, $p = .74$. The correlation coefficient between the two panels' weighted validity scores on the same country was $r = .90$ ($df = 14$, $p < .01$) and the dyadic (i.e. pairwise) intraclass correlation coefficient was $r_I = .90$ ($df = 30$, $p < .01$). The country-by-country validity weighted scores for both panels are shown graphically in Figure 1.

As illustrated in the figure, the Netherlands, New Zealand, United Kingdom, and United States models were the highest ranked in terms of their validity, although the Australian and Hong Kong models were not far behind, with the French model at the bottom. Major discrepancies in ratings on this metadimension were for Belgium (±24), Hungary (±16), and the Netherlands (±12).

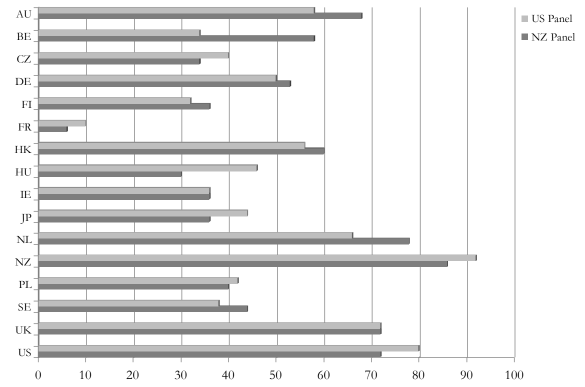

Figure 1. Profile of Validity Weighted Scores

Note. * The possible range of validity weighted scores was from 0-100, or 0%-100%.

Credibility

In addition to being valid, a good research evaluation model should produce conclusions that are believable or have reasonable grounds for being believable to relevant audiences; that is, a good model should be credible. The average difference between the two panels' weighted credibility scores was 1.18%; $t(15) = .59$, $p = .56$. The correlation coefficient between the two panels' weighted credibility scores on the same country was $r = .92$ ($df = 14$, $p < .01$) and the dyadic (i.e., pairwise) intraclass correlation coefficient was $r_I = .91$ ($df = 30$, $p < .01$). The country-by-country credibility weighted scores for both panels are shown graphically in Figure 2.

As illustrated in the figure, the Australian, Hong Kong, the Netherlands, New Zealand, United Kingdom, and United States models were clustered as the highest ranked in terms of their credibility, with the French model at the bottom. Major discrepancies in ratings on this metadimension were for Poland (±12) and the United Kingdom (±24).

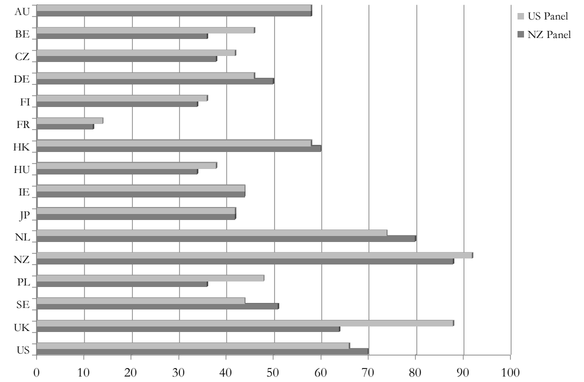

Figure 2. Profile of Credibility Weighted Scores

Note. * The possible range of credibility weighted scores was from 0-100, or 0%-100%.

Utility

In addition to being valid and credible, a good research evaluation model should be useful or designed for use; that is, a good model should have utility. The average difference between the two panels' weighted utility scores was 0.37%; $t(15) = .20$, $p = .84$. The correlation coefficient between the two panels' weighted utility scores on the same country was $r = .93$ ($df = 14$, $p < .01$) and the dyadic (i.e. pairwise) intraclass correlation coefficient was $r_I = .93$ ($df = 30$, $p < .01$). The country-by-country utility weighted scores for both panels are shown graphically and in tabular form in Figure 3.

As illustrated in the figure, the Hong Kong, the Netherlands, New Zealand, United Kingdom, and United States models were the highest ranked in terms of their utility, with the French model once again at the bottom. The only major discrepancy in ratings on this metadimension was for the Australian model (±18).

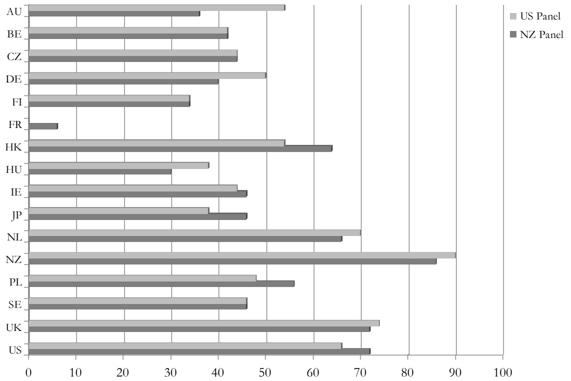

Figure 3. Profile of Utility Weighted Scores

Note. * The possible range of utility weighted scores was from 0-100, or 0%-100%.

Cost-Effectiveness

Besides having validity, credibility, and utility, a good research evaluation model should be economical in terms of the benefits produced by it; that is, a good model should be cost-effective. The average difference between the two panels' weighted cost-effectiveness scores was 1.37%; $t(15) = .52$, $p = .61$. The correlation coefficient between the two panels' weighted cost-effectiveness scores on the same country was $r = .84$ ($df = 14$, $p < .01$) and the dyadic (i.e., pairwise) intraclass correlation coefficient was $r_I = .83$ ($df = 30$, $p < .01$). The country-by-country cost-effectiveness weighted scores for both panels are shown graphically in Figure 4.

As illustrated in the figure, the Netherlands, New Zealand, United Kingdom, and United States models were the highest ranked in terms of their cost-effectiveness, although the Hong Kong model was not far behind. The French model is again at the bottom. Major discrepancies in ratings on this metadimension were for Hungary (±14) and the United Kingdom (±24).

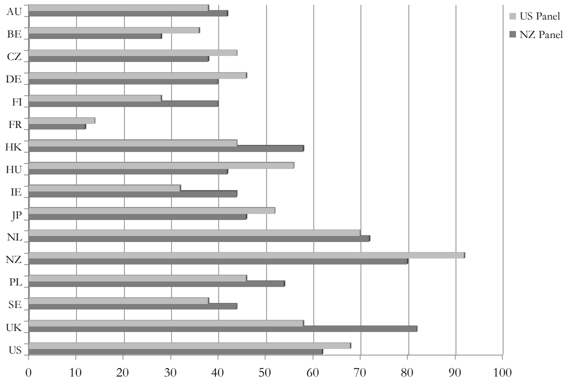

Figure 4. Profile of Cost-Effectiveness Weighted Scores

Note. * The possible range of cost-effectiveness weighted scores was from 0-100, or 0%-100%.

Ethicality

Finally, a good research evaluation model should be conducted in a legal, professional, and otherwise appropriate manner; that is, a good model should be ethical. The average difference between the two panels' weighted ethicality scores was 2.62%; $t(15) = .89$, $p = .39$. The correlation coefficient between the two panels' weighted ethicality scores on the same country was $r = .78$ ($df = 14$, $p < .01$) and the dyadic (i.e., pairwise) intraclass correlation coefficient was $r_I = .76$ ($df = 30$, $p < .01$). The country-by-country ethicality weighted scores for both panels are shown graphically in Figure 5.

As illustrated in the figure, the New Zealand, United Kingdom, and United States models were the highest ranked in terms of their ethicality. On this metadimension, the Netherlands national model was not rated as highly as on other dimensions, but still in the top five. Again, the French model is at the bottom. Major discrepancies in ratings on this metadimension were for Belgium (±28), Hungary (±18), and the United Kingdom (±18).

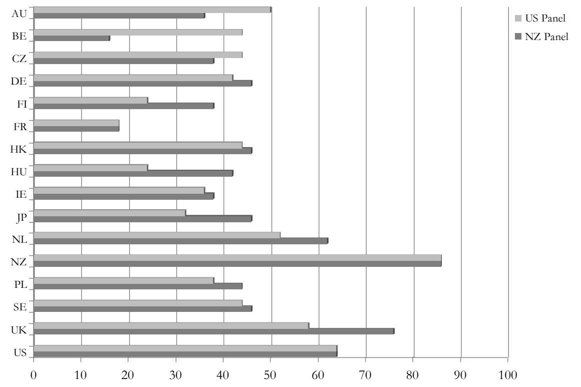

Figure 5. Profile of Ethicality Weighted Scores

Note. * The possible range of ethicality weighted scores was from 0-100, or 0%-100%.

Total Scores

The average difference between the two panels' total weighted scores was 0.65% (*t*[15] = .68, *p* = .51). The correlation coefficient between the panels' total weighted scores on the same country was $r = .98$ ($df = 14$, $p < .01$) and the dyadic (i.e., pairwise) intraclass correlation coefficient was $r_I = .98$ ($df = 30$, $p < .01$). The correlation coefficient between the rank order of total weighted scores was $r = .92$ ($df = 14$, $p < .01$), the dyadic (i.e., pairwise) intraclass correlation coefficient was $r_I = .92$ ($df = 30$, $p < .01$), and the Spearman's rank correlation coefficient was $\rho = .93$ ($df = 14$, $p < .01$). The country-by-country total weighted scores for both panels are shown graphically in Figure 6.

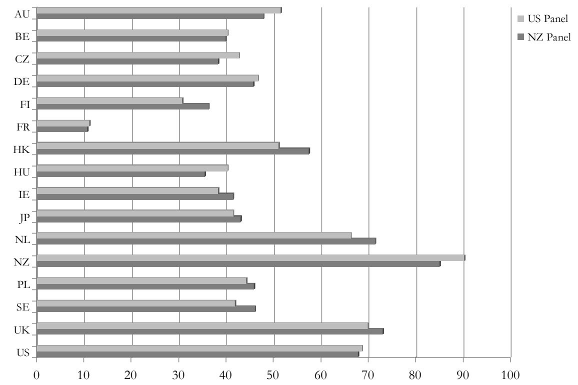

Figure 6. Profile of Total Weighted Scores

Concluding Remarks

In most countries, the competition for government research monies is getting increasingly competitive. This is particualrlly evident in systems that operate on performance-based funding (Coryn, 2007; Coryn, Hattie, Scriven, & Hartmann, 2007). Methodologically, large-scale research evaluations of government-financed research are most often binary in nature. That is, they are normally either a variant of traditional peer review (e.g., expert panels of one type or another) or are driven by indicators (e.g., student numbers, publications, external funding)—or, more often than not, a combination of the two. Both approaches have strengths and weaknesses. The indicator method, however, encourages the 'moral hazard;' that is, undue focus on productivity or assessment benchmarks, diverting attention away from more useful research into tactics for cultivating citations, for example.

As illustrated by the results of this study, a vast majority of national research evaluation models throughout the world can be characterized as less than ideal. However, most countries still regard their systems as experimental. On the whole, this study should provoke concern among policymakers and decision makers as there appear to be serious flaws and weaknesses in most nations' understanding and application of the reasoning and logic of evaluation.

Acknowledgements

This research was supported by the National Science Foundation, through the Division of Social and Economic Sciences (DSES), Decision, Risk and Management Sciences (DRMS) program (Award # SES-0721097). The authors would also like to thank the anonymous panel members (i.e., judges) for their contributions to this work.

References

Coryn, C. L. S. (2006). The fundamental characteristics of research. *Journal of Multidisciplinary Evaluation*, *5*, 124-133.

Coryn, C. L. S. (2007). *Evaluation of researchers and their research: toward making the implicit explicit.* Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and mechanisms for evaluating government-funded research: An international comparison. *American Journal of Evaluation*. In press.