Evaluation of Researchers and Their Research: Toward Making the Implicit Explicit

Chris L. S. Coryn
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
https://scholarworks.wmich.edu/dissertations/3401

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
EVALUATION OF RESEARCHERS AND THEIR RESEARCH: TOWARD MAKING THE IMPLICIT EXPLICIT

by

Chris L. S. Coryn

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation

Western Michigan University
Kalamazoo, Michigan
August 2007
Copyright by
Chris L. S. Coryn
2007
ACKNOWLEDGMENTS

A project of this magnitude is not a solitary endeavor and is possible only with the help of many, many people. First, and in addition to their contributions as individuals, I would like to begin by thanking my committee, Drs. Michael Scriven, John Hattie, and David Hartmann, for having the confidence in me to pursue this work. Thank you. The greatest contributor, and the one who deserves much of the recognition for this work, was Dr. Michael Scriven. He has been a tireless, selfless mentor and dedicated countless hours to this undertaking. Equally important were the contributions of Drs. Hattie and Hartmann; both of whom taught me to be a meticulous methodologist. Their support and encouragement made the process a truly rewarding experience. Thank you. It is not often that one has the opportunity to work with one scholar of such outstanding caliber, let alone three. Without them, I would never have made it through such a long and arduous process.

My first doctoral advisor and mentor, Dr. E. Jane Davidson, was a constant source of inspiration and without her diligence and dedication the Interdisciplinary PhD in Evaluation program would not even exist. Thank you. Dr. Catherine Borshuk of Indiana University, South Bend’s Department of Psychology also deserves special thanks for her truly outstanding mentoring, support, and kindness through both my undergraduate and graduate education. Thank you.

I would also like to extend my thanks to the current and former staff of The Evaluation Center at Western Michigan University for their support, who include Dr.
Acknowledgments—continued

Arlen Gullickson, Dr. Gary Miron, Dr. Carl Hanssen, Dr. Liesel Ritchie, Dr. Carolyn Sullins, Dr. Pamela Zeller, Dr. Daniel Stufflebeam, Dr. James Sanders, Mary Ramlow, Sally Veeder, and Christine Hummel. Without them I would never have made it through the Interdisciplinary PhD in Evaluation program. I sincerely thank you.

Considerable improvements were made to the dissertation due to the advice and insight of Dr. Lewis Pyenson, Dean of the Graduate College, Western Michigan University and Dr. Jennifer Wendling, Coordinator, Theses and Dissertations, Western Michigan University. Thanks.

Numerous researchers and evaluators throughout the world graciously donated their time and energy in reviewing and critiquing various aspects of this dissertation. Although they were “behind the scenes,” so to say, without them this dissertation would certainly have been much less than what it is and to them my sincere thanks and gratitude. These persons included: Dr. Claire Donovan, Research Evaluation and Policy Project, Australia; Dr. Linda Butler, Research Evaluation and Policy Project, Australia; Dr. Gretchen Jordan, Sandia National Laboratory, United States; Dr. Bill Valdez, Department of Energy, Office of Science, United States; Dr. Michael Holland, Chairman’s Designee, Energy Subcommittee Committee on Science, United States House of Representatives, United States; Dr. Dale Pahl, Assistant Laboratory Director for Research, Environmental Protection Agency, Office of Research and Development, United States; Dr. Dominic Orr, Hochschul-Informations-System GmbH, Germany; Dr. David Campbell, University of Klagenfurt, Austria; Dr. Shu-Tung Huang, Tung Nan Institute of Technology, Taiwan; Dr. Osamu Nakamura, Technology Evaluation and Research Division, Industrial Science and Technology Policy and Environment Bureau, METI; and the members of the American Evaluation Association’s Research,
Acknowledgments—continued

Technology, and Development Topical Interest Group and Washington Research Evaluation Network. Thanks should also be given to the European Commission for launching ERAWATCH—a research inventory covering most developed countries in the world—midway through the writing of this dissertation.

I also owe a great debt of appreciation to my truly outstanding colleagues and peers in the Interdisciplinary PhD in Evaluation program, including Daniela C. Schröter, P. Cristian Gugiu, Thomaz Chianca, Brandon Youker, Nadini Persaud, Amy Gullickson, Otto Gustafson, Willis Thomas, Ronald Visscher, Brad Watts, Paul Lamphear, John Risley, Wesley Martz, Ryo Sasaki, Mina Zadeh, Natasha Jankowski, Sheri Hudachek-Lantz, Todd Harcek, and Lori Wingate. From the School of Education at The University of Auckland, New Zealand, I would like to thank Dr. Earl Irving, Dr. Gavin Brown, Debbie Waayer, Dr. Susan Farrugia, Dr. Peter Keegan, and from the Auckland University of Technology, Dr. Richard Smith.

Also, I would like to thank the chapter contributors to the New Directions for Evaluation volume that emerged from this work—some of whom are mentioned elsewhere in this acknowledgement, but deserve additional recognition—including Dr. Michael Scriven, Dr. Claire Donovan, Dr. Gretchen Jordan, Dr. Jerald Hage, Dr. Jonathon Mote, Dr. David Newman, Dr. Kathleen Quinlan, Dr. Mary Kane, Dr. William Trochim, Dr. Bernd Ebersberger, Dr. Jakob Edler, Dr. Vivien Lo, Dr. Osamu Nakamura, Dr. Michiko Takagi Sawada, Dr. Shin Kosaka, Dr. Naoto Kobayashi, Dr. Denis Gray, Dr. Bride Mallon, Dr. Johnathan Morell, Dr. Connie Kubo Della-Piana, and Dr. Gabriel Della-Piana.

Finally, and most importantly, I want to thank my wife, Daniela Coryn Schröter, for her encouragement, support, love, and understanding, as well as her intellectual
contributions throughout writing this dissertation. Without her I suspect that this endeavor would never have been completed.

While this work is certainly imperfect and any significant omissions, mistakes, or oversimplifications are mine alone, they are nonetheless completely unintended. Finally, I must confess a certain hesitation in laying out these ideas because I have only a minimal training in science studies and even less in the history and philosophy of science. Hence, I want to be clear that what follows makes no attempt to be a definitive analysis.

This dissertation 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).

Chris L. S. Coryn
Due to its very nature, the evaluation of research permeates nearly every aspect of the work of researchers. They evaluate the work of others or have their own work evaluated. They evaluate hypotheses that come to mind, the previous literature, the quality of data, the explanatory power of theories, or the design of experiments or instruments. And this is not always casual evaluation. It is highly skilled evaluation, and becoming a first-rate or world-class researcher is a process of improving the quality of these evaluations. However, deciding when someone is or has become a first-rate or world-class researcher is an evaluation at a somewhat different level. It is a complex synthesis of judgments about how well the researcher does each of the constitutive types of evaluation, usually as evidenced in the work they are producing.

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.
The research in this dissertation describes, classifies, and comparatively evaluates the national models used to evaluate research and allocate research funding in sixteen countries. These models vary widely in terms of how research is evaluated and financed. However, nearly all share the common characteristic of relating funding to past performance. Each of these sixteen national models was rated on more than twenty-five quality indicators by independent, blinded panels of researchers and evaluators in two countries. The national models were then ranked in terms of their validity, credibility, utility, cost-effectiveness, and ethicality. The results of the rankings show that the clear leaders are nations using large-scale research assessment exercises of various hues. Bulk funding and indicator-driven models received substantially lower ratings. Implications for research evaluation practice and policy are considered and discussed.
PREFACE

The pursuit of knowledge 'for its own sake' can be traced back to the ancient Greeks, for whom this type of pursuit was associated with social status. Evaluation in research is going on almost all the time that a researcher works. He or she is evaluating hypotheses that come to mind, the previous literature, the quality of data, the design of experiments or instruments, for example. And this is not always casual evaluation. It is highly skilled evaluation and becoming a first-rate or world-class researcher is a process of improving the quality of these evaluations.

However, deciding when someone is or has become a first-rate or world-class researcher is an evaluation at a somewhat different level. It is a complex synthesis of judgments about how well the researcher does each of the constitutive types of evaluation, usually as evidenced in the work they are producing. In other words, it is not about components of scientific method, it is about the user of the methods.

In principle, the evaluation of research hardly needs elaborate philosophical underpinnings as there is normally, but not always, a consensus around what is truly important research. However, the conventional practices of evaluating researchers and their research have traditionally been implicit, subjective, and determined by a constitutive perspective of what constitutes good research. Making this process explicit, systematic, and objective requires a radical departure not only from the time-honored principles and procedures, which has predominately viewed science as a self-regulating endeavor, but also a departure from many contemporary ideologies.
This is no trivial matter given that governments around the world invest billions of dollars per annum in research initiatives and agendas. Of late, many of these governments have also started to invest equally large sums in assessing the quality of the research produced by these investments, the intellectual, social and economic impacts of research ventures, and in some cases their standing in the world’s research spectrum. Indeed, the livelihood of individual researchers, research groups, departments, programs, and entire institutions often swing in the balance.

This dissertation sets forth to: (a) review the increasingly-large literature concerning the principles and procedures used to evaluate researchers and their research, as well as their underlying ideologies; (b) explain their shortcomings; (c) make clear why there are sufficient reasons to justify changing them; (d) demonstrate that research can be evaluated systematically and objectively; and (e) propose cogent alternatives and/or improvements to existing principles and procedures.

Ultimately, this dissertation is about the next step upward in the long process of making the implicit explicit; that is, the search for standards of merit to be applied for the evaluation of researchers and their research.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>SECTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>PREFACE</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xvii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xxi</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Origins of Research Evaluation</td>
<td>2</td>
</tr>
<tr>
<td>From Ancient Practice</td>
<td>3</td>
</tr>
<tr>
<td>Medieval Guilds</td>
<td>4</td>
</tr>
<tr>
<td>Scientific Societies</td>
<td>5</td>
</tr>
<tr>
<td>The Emergence of Explicit Evaluations of Research</td>
<td>8</td>
</tr>
<tr>
<td>Toward a Discipline of Evaluation</td>
<td>20</td>
</tr>
<tr>
<td>The Turning Point</td>
<td>22</td>
</tr>
<tr>
<td>Evaluation Scholarship</td>
<td>23</td>
</tr>
<tr>
<td>Systematizing Research Evaluation</td>
<td>24</td>
</tr>
<tr>
<td>The Continuum of Complexity and the Governance of</td>
<td>26</td>
</tr>
<tr>
<td>Research</td>
<td></td>
</tr>
<tr>
<td>Retrospective</td>
<td>28</td>
</tr>
<tr>
<td>Logic of Evaluation</td>
<td>29</td>
</tr>
<tr>
<td>Definitions of Evaluation</td>
<td>29</td>
</tr>
</tbody>
</table>

viii
Table of Contents—continued

CHAPTER

Evaluative Claims ................................................................. 33
The Tripartite Taxonomy ..................................................... 35
Values and Valuing .............................................................. 37
Criteria versus Indicators .................................................. 44
Fundamental Operations .................................................... 45
Secondary Operations ....................................................... 48
Core Dimensions of Evaluation ........................................... 51
  Process Evaluation ........................................................... 52
  Outcomes Evaluation ....................................................... 54
  Costs Evaluation ............................................................. 56
  Comparative Evaluation .................................................. 58
  Generalizability Evaluation ............................................... 59
Fields of Evaluation .......................................................... 60
  Program Evaluation ......................................................... 61
  Personnel Evaluation ....................................................... 62
  Performance Evaluation .................................................. 63
  Product Evaluation ........................................................ 64
  Policy Evaluation .......................................................... 65
  Portfolio Evaluation ........................................................ 66
  Proposal Evaluation ....................................................... 67
The Transdisciplinary Model ................................................ 68
Table of Contents—continued

CHAPTER

Reasons and Motives for Evaluating Research ........................................... 72
Accountability and Efficiency ................................................................. 75
Resource Allocation ............................................................................. 76
Improvement ......................................................................................... 77
Synthesis ................................................................................................. 78
Decision Making ..................................................................................... 80
Norms and Standards of Research ......................................................... 82
Scientific Method .................................................................................. 82
Ethos of Science ................................................................................... 85
Methodological Rigor ............................................................................ 89
Properties of Good Research .................................................................. 95
Definitions of Research ........................................................................ 95
The Search for Criteria .......................................................................... 100
Implicit Criteria ..................................................................................... 102
Explicit Criteria ..................................................................................... 103
Conclusions .............................................................................................. 106

II. INTERNATIONAL RESEARCH EVALUATION MODELS ..................... 108
The International Research Landscape ................................................... 111
Historical Overviews, Research Structures and Contexts, and Evaluation Models ........................................................................ 114
Australia .................................................................................................. 115
Historical Overview .............................................................................. 115
Table of Contents—continued

CHAPTER

Evaluation Model................................................................. 165
Hong Kong............................................................................. 175
   Historical Overview...................................................... 175
   Research Structure and Context................................. 175
   Evaluation Model........................................................ 177
Hungary.................................................................................. 183
   Historical Overview...................................................... 183
   Research Structure and Context................................. 184
   Evaluation Model........................................................ 186
Ireland.................................................................................. 190
   Historical Overview...................................................... 190
   Research Structure and Context................................. 191
   Evaluation Model........................................................ 194
Japan..................................................................................... 199
   Historical Overview...................................................... 200
   Research Structure and Context................................. 200
   Evaluation Model........................................................ 202
Netherlands.......................................................................... 207
   Historical Overview...................................................... 207
   Research Structure and Context................................. 208
   Evaluation Model........................................................ 210
# Table of Contents—continued

## CHAPTER

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Model</td>
<td>165</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>175</td>
</tr>
<tr>
<td>Historical Overview</td>
<td>175</td>
</tr>
<tr>
<td>Research Structure and Context</td>
<td>175</td>
</tr>
<tr>
<td>Evaluation Model</td>
<td>177</td>
</tr>
<tr>
<td>Hungary</td>
<td>183</td>
</tr>
<tr>
<td>Historical Overview</td>
<td>183</td>
</tr>
<tr>
<td>Research Structure and Context</td>
<td>184</td>
</tr>
<tr>
<td>Evaluation Model</td>
<td>186</td>
</tr>
<tr>
<td>Ireland</td>
<td>190</td>
</tr>
<tr>
<td>Historical Overview</td>
<td>190</td>
</tr>
<tr>
<td>Research Structure and Context</td>
<td>191</td>
</tr>
<tr>
<td>Evaluation Model</td>
<td>194</td>
</tr>
<tr>
<td>Japan</td>
<td>199</td>
</tr>
<tr>
<td>Historical Overview</td>
<td>200</td>
</tr>
<tr>
<td>Research Structure and Context</td>
<td>200</td>
</tr>
<tr>
<td>Evaluation Model</td>
<td>202</td>
</tr>
<tr>
<td>Netherlands</td>
<td>207</td>
</tr>
<tr>
<td>Historical Overview</td>
<td>207</td>
</tr>
<tr>
<td>Research Structure and Context</td>
<td>208</td>
</tr>
<tr>
<td>Evaluation Model</td>
<td>210</td>
</tr>
<tr>
<td>Country</td>
<td>Historical Overview</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>New Zealand</td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td>224</td>
</tr>
<tr>
<td>Sweden</td>
<td>230</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>240</td>
</tr>
<tr>
<td>United States</td>
<td>248</td>
</tr>
<tr>
<td>Fundamental Characteristics of International Research Evaluation Models</td>
<td>260</td>
</tr>
</tbody>
</table>
Table of Contents—continued

CHAPTER

Primary Reasons and Motives ............................................................... 261
Basic Units of Assessment ................................................................. 263
Core Methods ....................................................................................... 265
Key Indicators and Criteria ................................................................. 268
Systemization and Consistency ............................................................. 270
Funding System Archetypes ................................................................. 272
Conclusions ........................................................................................ 274

III. METHODOLOGY FOR COMPARATIVE EVALUATION OF INTERNATIONAL RESEARCH EVALUATION MODELS ........................................... 275

Method ................................................................................................. 276
Study Design ......................................................................................... 276
Judges ..................................................................................................... 280
Setting and Materials ........................................................................... 283
Measures ............................................................................................... 286
  Validity .............................................................................................. 287
  Credibility ........................................................................................ 288
  Utility ................................................................................................ 288
  Cost-Effectiveness .......................................................................... 289
  Ethicality ........................................................................................... 289
  Reliability and Validity of Measures ................................................. 290
Procedure ............................................................................................. 291
  Stage I............................................................................................... 292
CHAPTER

Stage II ........................................................................................................... 294
Analytic Approach ......................................................................................... 295
Scoring ............................................................................................................. 296
Profiling ......................................................................................................... 298
Synthesis ......................................................................................................... 299
Human Judgment Analysis ............................................................................ 302
Summary ......................................................................................................... 303

IV. STAGE I AND II RESULTS: INDEPENDENT AND CONSENSUS RATINGS, SCORES, AND RANKINGS ................................................................................................................................. 304

Independent Judges’ and Panels’ Consensus Ratings .................................. 304
Primary Analyses and Results ........................................................................ 305
Validity Ratings, Scores, Rankings, and Profiles ........................................ 305
Credibility Ratings, Scores, Rankings, and Profiles ..................................... 308
Utility Ratings, Scores, Rankings, and Profiles .......................................... 311
Cost-Effectiveness Ratings, Rankings, and Profiles ................................... 314
Ethicality Ratings, Rankings, and Profiles .................................................... 317
Total Ratings, Scores, Rankings, and Profiles ............................................ 320
Quality Categories of National Models ......................................................... 324
Secondary Analyses and Results ................................................................. 326
Effectiveness of Bias Reduction via Blinding ............................................... 326
Reliability Estimates for Independent Judges and Panels ....................... 327
Reliability Estimates for Dyad-Pairs ............................................................. 332

xv
Table of Contents—continued

CHAPTER

Attributes Predicting Independent Judges' "Overall Best Judgment" ................................................................. 334

Generalizability of Judges' Ratings ............................................. 337

Classification of Funding System Archetypes ............................ 340

Summary of Key Findings .......................................................... 342

VII. CONCLUSIONS ........................................................................................................ 344

Central Findings .................................................................................. 344

Limitations ........................................................................................... 348

Implications ......................................................................................... 349

Theory and Practice ....................................................................... 349

Policy ................................................................................................. 350

Reforming the Evaluation of Research: Visions for the Future .......... 352

That Which can be Counted...Should it? ........................................ 353

The Research-Teaching Relationship Reconsidered .................... 353

The Role of Theory: A Cynical View ............................................ 355

Values Duality: The Tension between Competing Demands .......... 356

The Trifecta: Type A, B, and C Evaluations of Research ................ 358

Funding Research versus Evaluating Research: Getting it Right .............. 359

Peers...Collective Wisdom or Shared Bias? ................................. 360

The Holy Grail of Research Evaluation ........................................... 364

Concluding Remarks ....................................................................... 365
# APPENDICES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Acronyms</td>
<td>367</td>
</tr>
<tr>
<td>B. Judge Socio-Demographic Questionnaire</td>
<td>379</td>
</tr>
<tr>
<td>C. Judge Scoring Sheet</td>
<td>381</td>
</tr>
<tr>
<td>D. Judge Consent Form Version A</td>
<td>384</td>
</tr>
<tr>
<td>E. Judge Consent Form Version B</td>
<td>387</td>
</tr>
<tr>
<td>F. Human Subjects Institutional Review Board Approval</td>
<td>390</td>
</tr>
</tbody>
</table>

REFERENCES..................................................................................................................... 391
LIST OF TABLES

1. The International Research Landscape .......................................................... 113
2. Australia's Research Quality Framework Rating Scale .................................. 120
3. Australia's Research Quality Framework Research Impact Rating Scale .......... 120
4. Hong Kong's Research Assessment Exercise Scoring Schema ....................... 181
5. Netherlands's Standard Evaluation Protocol for Public Research Organizations Five-Point Scale ................................................................. 212
6. New Zealand's Quality Evaluation Component Weighting ......................... 220
7. New Zealand's Quality Evaluation Total Weighted Score Calculation ............ 221
8. New Zealand's Quality Evaluation Total Weighted Scores and Quality Categories .................................................................................................................. 221
9. New Zealand's Quality Evaluation Quality Categories .................................. 222
10. United Kingdom's Research Assessment Exercise Rating Scale .................. 246
11. PART Results to Date ...................................................................................... 260
12. International Research Evaluation Models' Primary Reasons and Motives for Evaluating Research .................................................................................................................. 262
13. International Research Evaluation Models' Basic Units of Assessment ............ 264
15. International Research Evaluation Models' Key Indicators and Criteria .......... 269
17. International Funding System Archetypes ...................................................... 273
18. Socio-Demographic Characteristics of Judges by Panel ............................... 282
19. Metadimension Weighted Score Calculation .................................................. 296
List of Tables—continued

20. Conversion of Weighted Scores to Quality Categories ........................................... 301
21. Quality Category Descriptions .................................................................................. 301
22. Validity Weighted Scores .......................................................................................... 307
23. Credibility Weighted Scores ....................................................................................... 310
24. Utility Weighted Scores .............................................................................................. 313
25. Cost-Effectiveness Weighted Scores ........................................................................ 316
26. Ethicality Weighted Scores ........................................................................................ 319
27. Total Weighted Scores ............................................................................................... 322
28. Quality Categories of National Models ..................................................................... 325
29. Pearson’s Product-Moment Correlation Coefficients for Indicator Ratings and ‘Overall Best Judgment’ and Rank Orderings of the Magnitude of the Pearson’s Product-Moment Correlation Coefficients ........................................... 328
30. Pearson’s Product-Moment Correlation Coefficients for Panels’ Metadimensional Raw Scores ............................................................................................................. 329
31. Total Weighted Scores and Rank Ordering of National Models ............................... 330
32. Cronbach’s Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Independent Judges’ Ratings ..................................................................................... 331
33. Cronbach’s Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Panels’ Consensus Ratings ..................................................................................... 331
34. Pairwise Intraclass Correlation Coefficient for Panel Dyad-Pairs’ Raw Scores .............................................................. 333
35. Cronbach’s Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Combined-Panels’ Consensus Ratings ...................................................................................... 334
36. Summary of Regression Analysis for Attributes Predicting NZ Independent Judges’ “Overall Best Judgment” .............................................................................................................. 335
37. Summary of Regression Analysis for Attributes Predicting US Independent Judges’ “Overall Best Judgment” .............................................................................................................. 336
<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>38.</td>
<td>Summary of Regression Analysis for Attributes Predicting Combined</td>
<td>336</td>
</tr>
<tr>
<td></td>
<td>Judges' &quot;Overall Best Judgment&quot;</td>
<td></td>
</tr>
<tr>
<td>39.</td>
<td>Comparison of Standardized Beta Weights for Attributes Predicting NZ,</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>US, and Combined Judges' &quot;Overall Best Judgment&quot;</td>
<td></td>
</tr>
<tr>
<td>40.</td>
<td>Summary ANOVA Table of Variance Component Sources used for Generalizability Study</td>
<td>338</td>
</tr>
<tr>
<td>41.</td>
<td>Variance Accounted for by Components</td>
<td>339</td>
</tr>
<tr>
<td>42.</td>
<td>Wilk's Lambda and Canonical Correlation Coefficients for the Three</td>
<td>340</td>
</tr>
<tr>
<td></td>
<td>Funding System Archetypes</td>
<td></td>
</tr>
<tr>
<td>43.</td>
<td>Means and Standard Deviations on the Five Metadimensions for the</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td>Three Funding System Archetypes</td>
<td></td>
</tr>
<tr>
<td>44.</td>
<td>Standardized Discriminant Function and Structure Coefficients</td>
<td>341</td>
</tr>
<tr>
<td>45.</td>
<td>Group Centroids</td>
<td>342</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1. Research Evaluation in Time: The Continuum of Complexity ........................................ 27
2. The Transdisciplinary Model .......................................................................................... 70
3. Spatial Map of a Discipline, Field of Evaluation, and Field of Application within the Transdisciplinary Model ....................................................................................................... 71
4. The Transdisciplinary Model and the House of Evaluation ........................................... 72
5. Sample Countries ............................................................................................................. 110
6. Procedural Design of Comparative Evaluation ................................................................ 295
7. Hypothetical Country Profile .......................................................................................... 298
8. Profile of Validity Weighted Scores ............................................................................... 306
9. Profile of Credibility Weighted Scores .......................................................................... 309
10. Profile of Utility Weighted Scores ................................................................................. 312
11. Profile of Cost-Effectiveness Weighted Scores ............................................................. 315
12. Profile of Ethicality Weighted Scores ............................................................................ 318
13. Profile of Total Weighted Scores .................................................................................. 321
14. Profile of NZ Panel Metadimensional and Total Ratings ............................................ 323
15. Profile of US Panel Metadimensional and Total Ratings ............................................. 323
CHAPTER I

INTRODUCTION

This chapter begins by presenting a brief historical account of evaluation in general and evaluation in and of research specifically. Following this account is a section devoted to the logic of evaluation, which is intended to demonstrate the reasoning on which this dissertation is founded and with which it positions itself. Also in this chapter are sections that describe the general purposes of research evaluation, norms and standards of research, and some of the demonstrable properties of good research.

The chapter is not intended to describe the historical evolution of the evaluation of researchers or their research in detail, nor is it a discourse on the true logic or nature of evaluation, nor an attempt to identify the only characteristics or properties which define good, valuable, or important researchers or research. Instead, it might be better characterized as a prelude to the chapters that follow. This brief historical overview provides some important clues to the general nature, principles, and practices of present-day efforts at evaluating researchers and their research. As will be seen, these modern explicit practices are not far removed from the age-old implicit practices. Nonetheless, the chapter begins from the most humble of beginnings—the implicit, intrascientific process of evaluating one's own research or the research of one's predecessors,

1 'Science,' 'scientific,' and 'research' as used throughout this dissertation are used in a very general sense and not restricted to the natural sciences, basic or applied research. They also include the social sciences as well as the arts and humanities, both of which have been vastly neglected in the evaluation of research. In part, this neglect can be attributed to the failure to adequately define research in a reasonable manner, and what does and does not constitute research has been widely disputed (see the Definitions of Research subsection of this chapter).
colleagues, or peers—and ends with some of the demonstrably relevant merit-defining values and criteria by which researchers and their research should be evaluated.

**Origins of Research Evaluation**

It has been claimed that the *explicit*, systematic evaluation of research is in its infancy, and has only emerged as a legitimate domain of evaluative inquiry within the past few decades (Martin, 1997; OECD, 1997). Nevertheless, *implicit* evaluation in research is a very old practice, and the former surely has its roots in the latter. In part, the practice of implicit evaluation in research can be inferred from written records dating to as early as the 2nd century BC. Amongst these are the writings of the Greek scientist and mathematician Archimedes, in one of which he described the way he discovered many of his geometrical results, which was clearly an introspective, implicit evaluative endeavor in research:

...certain things first became clear to me by a mechanical method, although they had to be proved by geometry afterwards because their investigation by the said method did not furnish an actual proof. But it is of course easier, when we have previously acquired, by the method, some knowledge of the questions, to supply the proof than it is to find it without any previous knowledge (Archimedes, 287-212 BC).

Evaluation is an essential characteristic of the human condition and is perhaps the single most important and sophisticated cognitive process in the repertoire of human reasoning and logic, and one which has defied adequate explanation for nearly two millennia (Osgood, Suci, & Tannenbaum, 1957; Scriven, 1991). Without such processes...

---

2 Evaluation *in* research is understood as *implicit evaluation* conducted by scientists and researchers, as part of 'intra-scientific' practice. By contrast, evaluation *of* research is understood as *explicit evaluation*, which is considered an 'extra-scientific' as well as 'intra-scientific' practice (Langfeldt, 2002). However, both can be pragmatically understood as simultaneously intrinsic and extrinsic enterprises.

3 Analysis and restoration of the *Palimpsest* from which this quotation comes is expected to be completed in 2007 and is the only copy of Archimedes' *On the Method of Mechanical Theorems* (The Walters Art Museum, 2006).

4 One of the most important sources of social psychological thinking about the cognitive processes of evaluation is Charles Osgood and his psychology of meaning (Osgood, Suci, & Tannenbaum, 1957).
there is simply no means for distinguishing the bad from the good, the worthwhile from
the worthless, or the significant from the insignificant. By some accounts evaluation as a
systematic process is more than 4,000 years old, of which some of the earliest examples
may be the personnel evaluations conducted by the Chinese and Egyptian dynasties
(Frechting, 2002; Lu & Xie, 2005).  

From Ancient Practice

Scriven (1991) bolstered this argument, in the introduction to the Evaluation
Thesaurus, by claiming that evaluation itself, and by the same token the evaluation in and
of research, is probably even older than the recently discovered Palimpsest of Archimedes
and the personnel evaluations of the Chinese and Egyptians:

Evaluation is a new discipline but an ancient practice. The earliest craft
workers of which we have a record, the stone-chippers, left a track record
of gradually improving quality of materials and design...there is no craft
without evaluation...in some crafts the evaluation activity has reached
considerable heights... (Scriven, 1991, p. 3).

Examples of this sort occur throughout history. For instance, the increasingly
sophisticated practices of the 1st century Japanese swordsmiths in forging, folding, and
hardening steel from iron and carbon, where nearly perfect swords were created with
simple tools, was no accident (Sato, 1983).  
It was a result of systematically applied
process and product evaluations of the swordsmiths' techniques, materials, and swords.

---

5 Although they had been taking place for several hundred years, the Chinese civil servant examinations
were perfected during the Tang dynasty (618-907 AD). This competitive procedure was designed to draw
the best talent into Chinese government. The practice of testing can also be traced to this period in China
(McDonald, 1999).

6 About 12,000 years ago there was a radical change in stone technology, which had been basically the
same for several million years. Instead of merely chipping and flaking stones to make tools, our ancestors
began to grind and polish them to make a wide variety of new tools (Stout, 2002; Toth, Clark, & Ligabue,

7 The art of weapon making was originally brought to Japan from China and Korea, roughly around the 1st
century. Although Chinese and Korean swordsmiths taught the Japanese how to make swords, it was the
Japanese who perfected the art of swordmaking (Sato, 1993).
One improvement resulting from these evaluations, around the 2nd century, was the insertion of softer cores into the blades which significantly heightened sword flexibility (Mishina, 1996). These new blades could withstand the impact of heavy blows on armor without shattering. Moreover, each sword made had to be worthy of being signed on the blade’s tang by the maker, and to justify claims about the quality of the blades all test results were also recorded on the tang (Kapp, Kapp, & Yoshihara, 2002; Sato, 1983).

**Medieval Guilds**

Similar practices occurred around the 11th century, with the creation of artisan’s and craftsmen’s guilds in Europe. These guilds were responsible for developing strict rules governing product and service quality (Epstein, 1991; Kieser, 1999; Lequin, 1986). Inspection committees enforced the rules by marking flawless goods with a special mark or symbol (Maguad, 2006; Wolek, 1999). Craftsmen themselves often placed a second mark on the goods they produced. At first this mark was used to trace the origin of faulty items, however, over time the mark came to represent a craftsman’s good reputation (Merges, 2004, November). For instance, stonemasons’ marks symbolized each guild member’s obligation to satisfy his customers and enhance the trade’s reputation (American Society for Quality, 2006). Inspection marks and master-craftsmen marks served as proof of quality for customers throughout medieval Europe (Kieser, 1999). In part, the quality of guilds’ products and services was maintained and improved

---

8 Regulated professions have existed for millennia, and, for example, The Code of Hammurabi specified death for builders or masons whose buildings collapsed on their inhabitants. The Hippocratic Oath applies to this day as the basis of the modern physicians’ ethical code. Furthermore, all known legal codes include some limits on the practices or powers of jurists, and it has generally been recognized that those in a position of special knowledge or trust were to be held accountable to the public for their advice and services (Harper, 1994/1904).

9 The artisan’s and craftsmen’s guilds ranged in nature from apothecaries to writers of text.

10 Many silversmiths still use these markings, which are often referred to as silver marks, hallmarks, and maker’s marks (Clifford, 1999).
by selecting the best apprentices, journeymen, and masters via a rigorous, highly competitive personnel evaluation system (Lequin, 1986; Smith, 1998).

Poverty, famine, and disease are often used to describe the Middle Ages, however, it was also during this period, around the 12th and 13th century, when there was a radical increase in the rate of new inventions and major scientific and technological advances (White, 1974). Some notable advances of the time included spectacles, artesian wells, the compass, the astrolabe, and dramatic improvements in ship making and clock making (White, 1975). During the early 14th century, the course of technology was forever changed with the introduction of gunpowder, which was claimed to have been discovered almost simultaneously in five different parts of the world (Kaempffert, 1941; Ruffell, 1992).

Scientific Societies

Several hundred years after the appearance of the European guilds, around 1660, when the Inquisition of the Holy Roman and Catholic Church was investigating ‘suspected novelties,’ (Feldhay, 1995; Langford, 1992) members of other secretive or informal societies already in existence formed the Royal Society of London for the Improvement of Natural Knowledge, simply referred to as the Royal Society (Purver &

---

11 Guild practices also included forms of product evaluation. For example, in order to become a master himself and join a guild, a journeyman had to demonstrate his skill in his craft by creating a masterpiece that was approved by the guild (American Society for Quality, 2006; Lequin, 1986).
12 It has been claimed that these advances led to the so-called Age of Discovery or Age of Exploration (Crone, 1961).
13 The discovery has been claimed by the Chinese, Hindus, Greeks, Arabs, Germans, and English (Kaempffert, 1941). However, “within the well-recorded histories of the first four there is no written evidence which would satisfy a historian that any of them discovered or used gunpowder before it came into use in Europe” (Ruffell, 1992). Fireworks of some sort are thought to have been in use as early as the 6th century in China, but these could have been made without gunpowder. Nevertheless, by the early 11th century various propellant mixtures containing saltpeter, sulfur, and charcoal were developed by the Chinese. Moreover, gunpowder is considered one of the the Four Great Inventions of ancient China (Tsou, 1998); known in Chinese as 四大发明 and in Pinyin as sì dà fāmíng. The other three are papermaking, the compass, and printing. These inventions are particularly celebrated in Chinese culture for their historical significance and as signs of ancient China’s advanced science and technology (Ronan, 1994).
The motto of the Royal Society, *Nullius in Verba*, signified the Royal Society’s commitment to establishing the truth of scientific matters through experiment rather than through citation of authority (Sprat, 2003). The Royal Society was dedicated to the free flow of scientific information and communication, and imagined a global network which would form the basis for an ‘empire of learning’ (Sorrenson, 1996) and only accepted the ‘best’ scientists of the time for membership, based primarily on their work. Notable founding members of the Royal Society included Robert Boyle, John Evelyn, Robert Hooke, William Petty, John Wallis, Thomas Browne, John Wilkins, Thomas Willis, and Sir Christopher Wren (Skinner, 1969).

Sir Isaac Newton, President of the Royal Society from 1703-1727, envisioned what is now referred to as modern ‘scientific method’ (1687/1999), which was considered fundamental to the investigation and acquisition of new knowledge based upon physical evidence (Purver & Bowen, 1960; Sprat, 2003). Although Newton formalized the scientific method, the rudiments of the scientific approach to knowledge can be observed throughout human history (Gower, 1997). Other early writers on scientific method included the English philosopher and statesman Francis Bacon (1620/1994), who wrote in the early 17th century that a tabulation of a sufficiently large number of observations of nature would lead to theories accounting for those

---

14 In 1542 a permanent congregation staffed with cardinals and other officials, established by Pope John III, was tasked with maintaining and defending the integrity of the faith and to examine and proscribe errors and false doctrines (Feldhay, 1995; Langford, 1992). In 1616 the congregation gave their assessment of the propositions that the Sun is immobile and at the center of the universe and that the Earth moves around it, judging both to be ‘foolish and absurd in philosophy,’ and the first to be ‘formally heretical’ and the second ‘at least erroneous in faith’ in theology (Armitage, 1951). This assessment led to Copernicus’s (1543/1992) *De Revolutionibus Orbium Coelestium* (On the Revolution of Heavenly Spheres) being placed on the Index of Forbidden Books (Gingerich, 2004), until revised, and Galileo Galilei to be admonished about his Copernicanism (Kuhn, 1957). In 1633 Galileo was tried and condemned for a grave suspicion of heresy, and all of his works were banned (Halsall, 1999).

15 Informal meetings had been taking place since the 1640s.

16 From Latin, meaning ‘on the words of no one.’

17 The scientific method is given greater attention in the *Norms and Standards of Research* section of this chapter.
operations; known as the method of *inductive reasoning* (Cohen, 1980). At about the same time, however, the French mathematician and philosopher René Descartes (1640/1968) was attempting to account for observed phenomena on the basis of what he called clear and distinct ideas; known as the method of *deductive reasoning* (Rips, 1994).

Another scientific society which profoundly influenced modern research and scientific thought was the Vienna Circle of Austria (Hanfling 1981; Jörgensen 1951; Kraft 1950/1968; Richardson 2003), simply referred to as the Circle, and organized under Moritz Schlick (Stadler, 2001). The Circle met weekly, for the most part, beginning in 1922 and ending in 1936, when Schlick was shot to death by an irate graduate student. By then many members had left Austria during the rise of the Nazi party due to clashes over its ideological and mysticism-based scientific research, and their approach to philosophy came to be known as *logical positivism*, which in part emerged from the Unity of Science conferences that took place between 1929 and 1941 (Ayer, 1958; Stadler, 2001).

Prominent members of the Circle included Rudolf Carnap, Otto Neurath, Herbert Feigl, Philipp Frank, Friedrich Waismann, Hans Hahn, and Ernst Topitsch (Kieseppä, 2002). They were visited on occasion by Hans Reichenbach, Kurt Gödel, Carl Hempel, Alfred Tarski, Willard Quine, and Alfred Ayer.18 Karl Popper, though he never attended the Circle’s meetings, was influential in the reception and criticism of their doctrines (Hacohen, 2001).

For some time a few of the Circle’s members met regularly with Ludwig Wittgenstein. However, while his *Tractatus Logico-Philosophicus* (1922/1999) was tremendously influential to the Circle, Wittgenstein came to feel that they had misinterpreted his work, and was frequently frustrated by these meetings (Finch, 2001; 18 Ayer popularized the Vienna Circle’s work in Britain and ultimately created his own *logical empiricism* (1958).
McGuinness, 2002). Partly as a result of his frustration with these meetings he was led to believe that there were grave errors in his work as presented in *Tractatus Logico-Philosophicus*, spurring him to a shift in his philosophical views, as would eventually be evidenced in his *Philosophical Investigations* (1953). However, logical positivism disappeared soon after, mainly due to Karl Popper (1934/2000a).

Scientific societies have existed in some form at least since the Greek Sophists of the 5th century BC (McClellen, 1985). From these groups there ultimately emerged, through the process of reviewing and selecting their members, explicit attempts at evaluations of the merits of researchers and their research. In more recent times, the evaluation of research has been dominated by variants of peer review and later by various quantitative approaches such as bibliometrics.

The Emergence of Explicit Evaluations of Research

...we are all quite familiar with Gregor Mendel (1822-1884) and his groundbreaking theories of heredity which paved the way to modern genetics. Unfortunately, Mendel's peers did not recognize his work at the time. Mendel sent his research report to the eminent Swiss botanist Karl Nägeli (1817-1891), who had developed his own theory of evolution known as orthogenesis. He therefore rejected Mendel's discovery of how heredity works in pea plants. Mendel wrote to Nägeli, summarizing his results and asking where it would be best to publish them. Nägeli answered that the experiments were worthless and should not be published at all. Charles Darwin also received a letter from Mendel but he did not even read it (Dinges, 2006, p. 11).

Explicit attempts at the evaluation of researchers and their research essentially emerged with the introduction of formalized peer review, which can only be traced back to the appearance of formal scientific journals in the 17th century (Langfeldt, 2002; Pyenson & Sheets-Pyenson, 1999). Most qualified experts (i.e., peers) were members of insular groups located in institutions of higher learning (i.e., universities). Although Leinster-Mackay (1978) traced the origins of the university back to the medieval period,
as places of teaching and advanced knowledge, modern universities only emerged in the 13th century, and as places of research, universities only began to gain prominence in the late 19th and early 20th century German universities (Hochschulen). Until then, many, in particular Cardinal John Newman (1853/1996), were adamant that research had no place in universities and that “the main aim of a university was teaching universal knowledge, whereas research was best undertaken outside the university” (Hattie & Marsh, 1996, p. 507).

As a cornerstone of modern scientific method, however, peer review has only been consistently applied since the middle of the 20th century (Aksnes, 2005). Before then its application was infrequent, and prior to World War II there were no universally adopted standards or norms for evaluating scientific research, and practices were conducted independently by each journal in response to idiosyncratic conditions (Burnham, 1992; Campanario, 1998). For instance, Albert Einstein's influential and groundbreaking *Annus Mirabilis* papers which appeared in the 1905 issue of *Annalen der Physik* were not peer-reviewed. In fact, the journal's editors, Max Planck and Wilhelm Wien—who were to later win the Nobel Prize in physics—simply recognized the virtue of such innovative ideas and published the papers without having them reviewed (Brown, 2005; Miller, 1981). Einstein's only genuine encounter with anonymous peer review occurred when he and Nathan Rosen, his first American assistant, submitted a paper to the *Physical Review* in 1936 on gravitational waves (Kenneflick, 2005). The paper received a negative review and Einstein reacted angrily to the referee's report:

> We (Mr. Rosen and I) had sent you our manuscript for publication and had not authorized you to show it to specialists before it is printed. I see

19 In his *Idea of a University*, Newman (1853/1996) contended that “to discover and to teach are distinct functions; they are also distinct gifts, and are not commonly found united in the same person” (p. 10).
20 Within that year Einstein published five papers (1905a, 1905b, 1905c, 1905d, 1905e), which later became “prime sources of three fundamental fields in physics: quantum theory (QT), theory of Brownian movement (BM), and theory of relativity (TR)” (Bushev, 2000, p. 380).
no reason to address the—in any case erroneous—comments of your anonymous expert. On the basis of this incident I prefer to publish the paper elsewhere (Kenneflick, 2005, p. 43).

In the post-World War II era, peer review practices have become increasingly more sophisticated and systematic with the introduction of double-blind and single-blind peer review (DBPR and SBPR) procedures—as opposed to open review—for controlling and assessing quality as well as disseminating research (Campanario, 1998). The double-blind process is one where not only the referees remain anonymous to the authors, but where the authors also remain anonymous to the referees, whereas single-blind procedures are where the reviewer knows the identity of the author but not vice versa (Justice, Cho, Winker, Berlin, & Rennie, 1998; Mainguy, Motamedi, & Mietchen, 2005; McNutt, Evans, Fletcher, & Fletcher, 1990). In open peer review, the identities of both authors and reviewers are revealed, affording the authors the ability to identify the reviewers (van Rooyen, Godlee, Evans, Black, & Smith, 1999; Walsh, Rooney, Appleby, & Wilkinson, 2000). The main argument against open peer review is that junior reviewers will be reluctant to criticize the work of senior researchers for fear of reprisals. This fear is particularly acute for researchers whose livelihoods depend on winning grants (Smith, 1999). The principal argument in favor of blinding is that:

…the signing of reviews would inhibit reviewers from being open and probing in their critiques (as has increasingly happened with letters of personal recommendation); this would clearly not be in the best interests of good science. The principal argument against blinding is that it might foster irresponsibility, particularly slanted and destructive criticism, because reviewers know that authors cannot hold them personally accountable for their opinions. The case for “opening up” peer review by identifying reviewers to authors is therefore being vigorously put forward (Davidoff, 1998, p. 66).

Modern peer review is almost universally the predominant method used for evaluating research, by and large seen as an obligatory system within the scientific community, and widely perceived as the only legitimate method for valuing scientific
merit (Campanario, 1998; Davidoff, 1999; Langfeldt, 2002; Smith, 1998). Nonetheless, it has been heavily attacked. For example, it is claimed that peer review is “partial, biased and unreliable, and it takes time away from research activities” (Langfeldt 2002, p. 16). It is, more or less directly, claimed that the peer review system is essentially an ‘old boys’ club,’ that it is full of scientists feathering their own nests, that it favors eminent scientists, that it stifles innovative research because assessments are done by well-established researchers rejecting ideas which differ from their own, that it discriminates against scientists who work in ‘low-prestige’ institutions, that it is slow, that it is expensive, that it is punitive against innovation, sometimes incompetent, that it suffers from the tyranny of small numbers, and that it is often unable to detect fraud (Boaz & Ashby, 2003; Cole, Cole, & Simon, 1980; Grayson, 2002; Langfeldt, 2002; McCook, 2006; Schröter, Coryn, & Montrosse, 2006; Turney 1990). Too often, discussions of peer review narrowly focus on technical matters such as inter-rater agreement, conflicts of interest, and normalization of raters’ scores to achieve comparability across panels (Hackett, 1997); there are, however, additional concerns which will be discussed later.

Peer review has traditionally been a process used to evaluate individual researchers or research products for decisions about employment, promotion, publication, awards, and funding of research projects. However, in the last twenty years, the large-scale use of peer panels to evaluate larger units, such as research groups, research disciplines, and institutes and research programs, in order to allocate resources and set research policy priorities, has become commonplace (ab Iorwerth, 2005; 21 Modern forms of peer review are premised on the assumption “that a judgment about certain aspects of science, for example its quality, is an expert decision capable of being made only by those who are sufficiently knowledgeable about the cognitive development of the field, its research agendas and the practitioners within it” (OECD, 1987, p. 28).

22 This is often referred to as the halo effect (Asch, 1946; Thorndike, 1920).

23 That is, a small number of reviewers are not a ‘representative’ sample of scientific opinion.

24 A 2005 poll of 3,247 scientists (Martinson, Anderson, & de Vries, 2005) funded by the National Institutes of Health (NIH) found 0.3% admitted faking data, 1.4% admitted plagiarism, and 4.7% admitted to autoplagiarism (i.e., republishing the same material or data).
In addition to scientific merit, these large-scale panels are often concerned with the working conditions of researchers, the socio-economic impact of research, or the utilization of research results, for example.

The term peer review is now reserved only for the more traditional review and assessment systems of scholarly communities such as reviews of manuscripts for scholarly journals, review of applications for academic positions (including appointments, promotions, demotions, and so forth), and review of grant applications. The term used for evaluations of research conducted beyond the individual instance or piece of research or individual researcher, however, is expert panel evaluation, which can be of two types: peer panel evaluation, and mixed panel evaluation (Langfeldt, 2002). Expert panel evaluations are those commissioned, often ad hoc, for evaluations at the program, institutional, or discipline level. Peer panel evaluation is expert panel evaluation when the panel consists of researchers qualified in the area under review. When the expert panel evaluation consists of both peers and other experts (e.g., experts on policy or commercialization of research) it is referred to as mixed panel evaluation.

Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy. In contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants) (Langfeldt, 2002, p. 18).

Fundamentally, both types are aimed at the allocation of scarce resources; the former in the terms of journal space or funding, and the latter also for funding, but in addition to set research priorities, to identify strengths and weaknesses in research areas,

---

25 These systems are given greater attention in Chapter II.
and so on. Both, however, are also used to control research communities and it is a zero sum game. Evaluatees are graded and ranked, and losers and winners are identified, especially by expert panel evaluations. There are, however, important differences. For example, peer review is mostly anonymous and confidential, whereas expert panel evaluations are normally public and reviewers identifiable.

By and large, the underlying ideal behind both peer review and expert panel evaluation is that of objectivity and the triangulation of judgments (more so if they form their conclusions before convening). Some efforts have been made to use quantitative measures in this process, for example, beginning in about 1945, journal editors began to use null hypothesis significance testing (NHST)\textsuperscript{26} results to determine which research studies to publish, “respectively, those with or without statistically significant results...to remove subjective judgment” (Kline, 2004, p. 8). Despite its importance, and with all its strengths and its flaws, the peer review process itself has until recently not been subject to the same intense scrutiny as the content of scholarly work.\textsuperscript{27}

Objectivity has been a normative notion of the scientific endeavor almost as long as modern science has existed (Longino, 1990; Porter, 1995). This notion, however, has never been taken to a greater extreme than Max Weber’s \textit{value-free doctrine}, which was introduced in the early 1900s (Scriven, 1991).\textsuperscript{28} The value-free doctrine is premised on the view that scientists were supposed to work without value-judgments, that is

\begin{itemize}
\item \textsuperscript{26} Acronyms used throughout this dissertation can be found in Appendix A.
\item \textsuperscript{27} To date, the most complete and critical analysis of the peer review system has been Cicchetti’s (1991) review of the reliability of peer review. Not surprisingly, he found that the reliability of most reviews is even less than would have occurred by chance.
\item \textsuperscript{28} The value-free nature of science, as proposed by Weber, claims that statements of fact are one thing, statements of value another, and any confusing of the two is impermissible (1904/1949). Although Weber is most often credited with the objection to values in science (1904/1949), he was actually preceded by more than a century by David Hume (1740/1985) and followed thereafter by George Moore (1903/1993) and his \textit{naturalistic fallacy}.
\end{itemize}
objectively or value-neutral (Weber, 1904/1949). But the evaluation of research is not value-free.

Much of the study of the history of science has been devoted to answering questions about what science is, how it functions, and whether it exhibits large-scale patterns and trends (Lakatos, 1971). The sociology of science has focused on the ways in which scientists work, examining the ways in which scientific knowledge is produced and constructed (Curd & Cover, 1998). Since the 1960s, a common theme in the study of the sociology and history of science has been to emphasize the human element of scientific knowledge, and to deemphasize the view that scientific data is self-evident, value-free, and context-free (Brillouin, 1962/2004).

A major subject of concern and controversy in the philosophy of science has been the inquiry about the nature of **theory change** in science (Darden, 1991). Three philosophers in particular who represent the primary poles in this debate have been Karl Popper (1963), who argued that scientific knowledge is progressive and cumulative; Thomas Kuhn (1962/1996), who argued that scientific knowledge moves through paradigm shifts and is not necessarily progressive; and Paul Feyerabend (1975), who argued that scientific knowledge is not cumulative or progressive, and that there can be no demarcation between science and any other form of systematic investigation.

Since the publication of Kuhn's *The Structure of Scientific Revolutions* in 1962, there has been much debate in the academic community over the meaning and objectivity of science. Often, but not always, a conflict over the truth of science has been split along the lines of those in the scientific community and those in the social sciences or humanities (Hollis, 1994). Part of the problem with Kuhn's conception of scientific revolutions is that scientific paradigms supposedly define truth, in the sense of the

---

29 According to Weber (1904/1949) values are personal judgments or preferences about what is considered good or bad, or about what is liked or disliked.
current set of beliefs in which scientists operate. They certainly define *prima facie* truth, “but that’s a long way from being the same; eventually paradigms are rejected as being too far from reality and they are always governed by that possibility” (Scriven, 1991, p. 253). The natural progression in the quest for objectivity and truth, of the merits of researchers and their research, has led to attempts at quantifying evaluations of research, in part due to the growing skepticism towards the objectivity of peer review.

...I am continually interested in the question of how we know whether or not someone is ‘good’ or ‘excellent’ in their field. The easiest way is to ask how the field how it discriminates...‘internationally referred research’ publications (which usually means publication in American journals—Americans oddly, don’t insist on publication in Australian journals!) make up only one measure...the central point is that we need to discriminate between the good and the less good, and in the domain of research and scholarship that is not an easy thing to do (Aitkin, 2000, pp. 151-152).

Quantitative evaluations of research have generally been conducted by scientometricians, bibliometricians, and information and library scientists using indicators of *quantity, quality, impact, or influence* (Coryn, 2006a; Research Evaluation and Policy Project, 2005). Bibliometrics is not a new phenomenon, and some of its earliest applications can be traced to the second half of the 19th century when Frank Shepard created a citation index—*Shepard’s Citations*—covering judicial decisions for attorneys to use for determining whether a legal procedure was still valid (Aksnes, 2005; Wouters, 1999). *Shepard’s Citations* had been in use as a legal reference tool since 1873 and owed its existence to the fact that “American law, like English law, operates under the doctrine of *Stare Decisis*” (Weinstock, 1971, p. 188).30 In addition, some research librarians had systematically applied bibliometric analyses since the early years of the 20th century. Among the first examples of such applications were studies of the frequency of journal

---

30 *Stare Decisis* means that all courts must follow their own precedents as well as those established by higher courts (Weinstock, 1971). Before presenting the previous decision as a precedent, however, attorneys must make sure that the decision had not been overruled, reversed, or limited in some way. *Shepard’s Citations* made this possible with a minimum of difficulty.
citations (Cole & Eales, 1917), which were considered to indicate the value of subscriptions to journals (Gross & Gross, 1927).

The genesis of modern bibliometrics, however, is primarily based on the work of Derek de Solla Price and Eugene Garfield. Price was a science historian and information scientist who is often credited as the father of scientometrics (Aksnes, 2005), whose work included the establishment of scientometrics31 as an independent discipline through his study of the exponential growth of science and the half-life of scientific literature (Price, 1963) and the examination of interactive communication patterns of scientists (Price, 1965). Garfield, however, is credited with the creation of the *Science Citation Index* (SCI), at the Institute for Scientific Information (ISI)32 in 1961, the impetus for modern bibliometrics (Glänzel, 2002a, 2002b).

This bibliographic database was originally developed for information retrieval purposes, much like Shepard's Citations, to aid researchers in locating papers of interest in the vast research literature archives. As a subsidiary function, however, it enabled scientific literature to be analyzed quantitatively, and has since become one of the mainstays for evaluating scientific research. Since the 1960s, the SCI and similar bibliographic databases have been applied in a large number of studies and in a variety of fields. Garfield (1979), in the introduction to his seminal book on the subject, *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*, stated that the rationale for this approach to evaluating research was that:

---

31 Scientometricians have become specialized in rating and mapping the natural sciences, the social sciences, and the arts and humanities with the help of huge databases derived from the scientific literature, usually on behalf of policy makers. Currently, however, there are only a few hundred scientometricians in the world. They vary from a lone individual who is part of a research library or history of science department, to a large collective with around twenty full-time researchers. Moreover, since the advent of computer mediated communication, which is rapidly becoming the principal medium for publication and dissemination of professional and scientific results, these changes may very well lead to the death of the discipline of scientometrics in its current form (Wouters, 1999).

32 Although the Institute for Scientific Information was recently renamed Thompson Scientific, ISI is used throughout this dissertation as this is the more widely known acronym.
Almost all papers, notes, reviews, corrections and correspondence published in scientific journals contain citations. They cite—generally by title, author and where and when published—documents that support, provide evidence for, illustrate, or elaborate on what the author has to say. Citations are the formal, explicit linkages between papers that have particular points in common. It [a citation index] lists publications that have been cited and identifies the sources of the citations (Garfield, 1979, p. 1).

Although Garfield's (1979) raison d'être appears commonsensical, there are demurrers, however, as Paul Wourters noted in The Citation Culture (1999):

...citing behaviour seems to vary according to personal traits. Whereas one author will devote detailed attention to the list of references, another could not be less interested (though this cannot be said too loudly)...[and]...the overall citing properties of the publications within a certain field share the same characteristics. The mathematician tends not to cite many publications. The biomedical researcher, on the other hand, is not afraid to cite hundreds of articles. The historian also likes references, but in a different way. The literary scholar goes about citing in quite another way...the sciences host many types of citing culture, each slightly different from the other...a scientist is supposed to cite honestly [italics added]; he must have read the article and have found it useful in some way. The question is, however, in what way this differs from the generally accepted norm of honesty. The moment one tries to become more concrete, and asks what it means to cite honestly and correctly, the answer becomes specialty bound (Wourters, 1999, p. 3).

Quantitative approaches aimed at assessment of research quantity, quality, impact, or influence can be grouped into two general classes: bibliometric and non-bibliometric (Coryn, 2006a; Research Evaluation and Policy Project, 2005). Bibliometric indicators are based on published literature in all of its forms, including, but not limited to, journal articles, monographs, books, book chapters, conference papers and proceedings, patents, and the references these publications contain. Bibliometric indicators are further divided into three sub-classes: publications; citations; and structural. Publication indicators include simple counts of publications and are usually viewed as a measure of research productivity, or quantity, rather than quality (Moed, 2005; Research Evaluation and Policy Project, 2005). Citation indicators are the references to earlier contributions upon
which a scientific work was built, and against which it positions itself, and assumes that
the “number of citations can be regarded as a measure of scientific quality or impact”
(Aksnes, 2005, p. 7). Structural indicators are auxiliary or proxy indicators which provide
information about the characteristics of research undertaken, such as publication
strategies or the place of a researcher or research unit in the scientific community, and
not normally considered indicators of research performance (van Raan & van Leeuwen,
2002).

*Non-bibliometric indicators* encompass all other quantifiable indicators such as
external funding and measures of esteem (e.g., honors and awards, editorship of journals,
and keynote addresses). A third type of quantitative indicator, belonging to the
bibliometric class, is the rapidly expanding web-link analysis, or *webometrics* (Bauer &

Advocates of quantitative approaches to the evaluation of research are at odds
on the role that quantitative indicators should play. Whereas some (e.g., Oppenhiem,
1997) suggest that quantitative evaluation of research should replace peer review, others
(e.g., ABRC, 1990; Aksnes, 2005; Aksnes & Taxt, 2004; van Raan & Leeuwen, 2004)
recommend that the role of quantitative evaluation of research is not to replace peer
review, but rather to enhance it.33

In part, the advocates’ case for replacing peer review with quantitative indicators
is based on studies that have found a positive relationship between peer review ratings
and various quantitative indicators. Among these are Luukkonen’s (1991) findings of a
tendency for citation counts to correlate roughly with peer ratings, Oppenhiem’s (1997)
study which found strong, positive correlations between citation indicators and the
United Kingdom’s Research Assessment Exercise (RAE) ratings for British research in

---

33 The use of quantitative indicators in the peer review process is often referred to as ‘informed peer
review.’
genetics, anatomy, and archaeology, and Rinia, van Leeuwen, van Vuren, and van Raan's (1998) study which found that various citation indicators correlated significantly with peer ratings of research programs in condensed matter physics in the Netherlands. Oppenhiem (1997), for example, argues in support of replacing peer review on the basis of the costs of the RAE peer judgment process, whereas others have a tendency to “point to bibliometric indicators as objective measures in contrast to the subjective character of peer review” (Research Evaluation and Policy Project, 2005, p. 5). Aksnes and Taxt (2004), however, found only weak correlations between bibliometric indicators and ratings given by expert committees in the case of Norwegian research groups, and Aksnes (2005) cautions that the fundamental distinctions underlying such comparative studies have gone largely ignored:

...peer review is oriented towards a decision context: is this paper eligible for publication, is this proposal eligible for funding? Aspects related to quality might well be one of the considerations in formulating the advice, but is not the only consideration...a peer-evaluation may involve assessments of factors that are not likely to be reflected through citation counts. Only when citation indicators are used in the same decision context as peer review and the two focus on the same aspect of the scientific performance can one reasonably compare them. Secondly, judgments made by peers may not necessarily be considered as the “truth” to which bibliometric indicators should correspond...the peers may be biased or mistaken in their assessments, or they may not be competent to judge...it is therefore a question of the extent to which peer assessments and citation indicators can be compared and be expected to correlate (Aksnes, 2005, p. 46).

Be that as it may, one of the leading proponents of bibliometric studies of research, Anthony F. J. van Raan (2005), asserts that:

...bibliometric assessment of research performance is based on one central assumption: scientists who have to say something important, do publish their findings vigorously in the open, international journal ('serial') literature. This assumption introduces unavoidably a 'bibliometrically limited view of a complex reality'. For instance, journal articles are not in all fields the main carrier of scientific knowledge; they are not 'equivalent' elements in the scientific process, they differ widely in
importance; and they are challenged as the 'gold standard' by new types of publication behaviour, particularly electronic publishing. However, the daily practice of scientific research shows that inspired scientists in most cases, and particularly in the natural sciences and medical research fields, 'go' for publication in the better and...if possible...the best journals. A similar situation is developing in the social and behavioural sciences, engineering and, to a lesser extent, in the humanities (van Raan, 2005, p. 134).

Perhaps equally important, however, as van Raan (2005) also pointed out, is the relatively recent proliferation of rapid and particularly cheap evaluations with the help of journal impact factors, or 'amateur' bibliometrics. For example, many departments merely count the number of citations to a researcher's publications or make the faulty assumption that publication in low impact journals are of low quality, or reciprocally that publications in high impact journals are of high quality, when reviewing faculty for promotions or tenure.

Following the introduction of formalized peer review systems and coinciding with the development of bibliometric and similar techniques was the transition of evaluation from a mere practice to a profession with independent disciplinary status (as indicated by increasing numbers of professional associations and societies, scholarly journals, and programs of study devoted to it), as well as the systemization of the evaluation of research. The next section of this chapter briefly addresses these changes and their impact on the evaluation of research.

Toward a Discipline of Evaluation

The origin of evaluation as a disciplined field of study has a relatively short history in contrast to informal or implicit versions of it. Evaluation of the professional...
variety, particularly program evaluation, has its roots in the United States’ ‘War on Poverty’ programs initiated in the 1960s under the Johnson and Kennedy administrations in the quest for a ‘Great Society.’ During the period from the 1950s through the 1970s, government spending for social programs in health, education, and housing, among others, increased from approximately $23 million in 1950 to more than $400 million by 1979 (Shadish & Luellen, 2005). Beginning in the 1960s, evaluation grew and flourished as a professional practice and the impetus for conducting evaluation was that:

...there is not enough money to do all the things that need doing...even if there were enough money, it takes more than money to solve complex human and social problems...not everything can be done, there must be a basis for deciding which things are worth doing (Patton, 1997, p. 11).

However, by the end of the 1960s, it was clear that evaluations of the Great Society programs had generally gone unnoticed or were overly politicized.

The utopian hopes for a scientific and rational society had somehow failed to be realized. The landing of the first human on the moon came and went, but poverty persisted despite the 1960s “War” on it—and research was still not being used as the basis for government decision making (Patton, 1997, p. 7).

Despite the poor utilization of many early efforts, evaluation continued to develop and the telltale signs of the transition of evaluation from a mere practice to a profession, and ultimately an independent discipline,35 were evident in the creation of professional publications, societies, codes of conduct,36 and training37 dedicated

35 Indeed, evaluation was not taken seriously as an academic discipline until the last third of the 20th century.

36 The first of which were the principles jointly prepared by the Evaluation Network (ENet) and the Evaluation Research Society (ERS) in the mid-1980s when the two organizations were considering a merger.

37 Presently, in the United States, graduate-level programs in evaluation are offered at: Boston College; Claremont Graduate University; Florida State University; George Mason University; Southern Illinois University, Carbondale; University of California, Berkeley; University of Connecticut; University of Illinois, Urbana-Champaign; University of Kentucky; University of Maryland; University of Minnesota; University of North Carolina; University of Texas, Austin; University of Wisconsin, Stout; Utah State University; and Western Michigan University (American Evaluation Association, 2006a).
exclusively to it. Not all agree that evaluation has completed the transformation from a practice to a discipline, however.

...[a] discipline evaluation is not. Disciplines are systematic, coherent, founded more often than not on sound theory, and offered as programs in accredited colleges, universities, and professional schools. Evaluation without detracting in the least from its multitude of contributions and creative authors and practitioners, is not systematic, coherent, theory-driven—oh perhaps with an exception here and there—as a program of study at institutions of higher learning. Evaluation is a helter-skelter mishmash, a stew of hit-or-miss procedures, notwithstanding the fact that it is a stew that has produced useful studies and results in a variety of fields...(Perloff, 1993 as cited in Davidson, 2005a).

By contrast, Eleanor Chelimsky (1997a) asserted that “evaluation is now spreading to the far corners of the globe...[and]...was adopted in 1992 as a key work area for the national auditing agencies of 188 independent states worldwide” (p. 54). Moreover, it had:

...routinely produced methodologically strong evaluations...timely enough and policy relevant enough to have practical value to decision makers...[and]...developed (through invention, or through begging, borrowing, and stealing) an extraordinary wealth of methods, styles, and reporting formats... (Chelimsky, 1997a, p. 53).

The Turning Point

One of the major turning points in the history of evaluation, and in its transition from practice to discipline, was the establishment of professional evaluation societies and associations. Among the earliest known professional societies were the May 12th Group (Madaus & Stufflebeam, 2000), Division H of the American Educational Research Association (AERA), the Evaluation Network (ENet),38 and the Evaluation Research Society (ERS), all of which developed in the early 1970s. ENet was founded by a five-member, self-appointed committee consisting of Daniel Stufflebeam, Egon Guba, Bill

38 ENet was originally funded, in part, by Phi Delta Kappa.
Gephart, Malcolm Provus, and Tom Hastings (L. Wingate, personal communication, March 16, 2006). Michael Scriven was ENet’s first elected President.

Originally, ENet and ERS existed independently, in part due to perceptions of some ERS members that ENet was a lesser organization because it consisted primarily of persons from the education sector. However, ENet and ERS merged in 1986, following the financial collapse of ERS\(^\text{39}\) to form what is presently known as the American Evaluation Association (AEA). AEA presently has more than 4,000 members representing “all 50 states in the United States, as well as 50 foreign countries” (Shadish & Luellen, 2005, p. 186). In 1995, there were only five evaluation organizations worldwide, including AEA, the Canadian Evaluation Society (CES), the Australasian Evaluation Society (AES), the European Evaluation Society (EES), and the Central American Evaluation Society (ACE), most in developed countries. By 2003 there were more than 50 national and regional evaluation organizations, most in developing countries (Chianca, 2004, January).\(^\text{40}\)

Evaluation Scholarship

There are now more than a dozen journals dedicated exclusively to evaluation scholarship, many of them published internationally. Evaluation and Program Planning, one of the discipline’s earliest publications, first appeared in 1974 (J. Morell, personal communication, February 17, 2006), followed by the publication of Evaluation Review in 1976, some years later by the American Journal of Evaluation (formerly published under the titles Evaluation News, prior to 1986, and Evaluation Practice, between 1986 and 1997), Evaluation Review: A Journal of Applied Social Research which was first published in 1976,

\(^{39}\)The collapse of ERS has been attributed in part to its attempt to publish a scholarly journal, which it could not afford (M. Scriven, personal communication, March 8, 2006).

\(^{40}\)Evaluation organizations were also recently established in Nicaragua and El Salvador (T. Chianca, personal communication, March 6, 2006), as well as New Zealand (aneza, 2006).
New Directions for Evaluation (formerly New Directions for Program Evaluation) and Evaluation & The Health Professions both of which appeared in 1978, Educational Evaluation and Policy Analysis which first appeared in 1979, the Canadian Journal of Evaluation which emerged in 1986, the Journal of Personnel Evaluation in Education which was first published in 1987, Practical Assessment, Research and Evaluation which was launched in 1988, Evaluation: The International Journal of Theory, Research and Practice which was first published in 1995 (in the United Kingdom), the Evaluation Journal of Australasia which was first published in 2000, and the Journal of MultiDisciplinary Evaluation which first appeared in 2004.

The journal Research Evaluation, devoted specifically to the evaluation of research, researchers, research groups, and scientific fields emerged only in the early 1990s, and is published in the Netherlands. This list is not exhaustive and does not include, for example, the numerous newsletters and other types of publications and scholarly communication, such as the Harvard Family Research Project's Evaluation Exchange which was first published in 1995.

Systematizing Research Evaluation

There was much hoopla about the rationality that social sciences would bring to the untidy world of government. It would provide hard data for planning...and give cause-and-effect theories for policy making...it would bring to the assessment of alternative policies a knowledge of relative costs and benefits so that decision makers could select the options with the highest payoffs...[and]...once in operation...their effectiveness...(Weiss, 1977, p. 4).

The earliest examples of the systemization of the evaluation of research has its roots in the large-scale evaluations of research conducted in Sweden in the late 1970s, which were shortly followed by similar evaluations of research in Finland, Norway, and Denmark in the mid-1980s (Luukkonen, 2002). AEA's Research, Technology and Development (RTD) Topical Interest Group (TIG) was established in 1995, partly in
response to increasing interest in evaluating federally-funded research in the United States. In 2001, the World Research Evaluation Network (ResEval)\(^{41}\) was established and followed in 2003 by the Washington Research Evaluation Network (WREN).\(^{42}\) ResEval was developed by the Georgia Institute of Technology through a National Science Foundation (NSF) sponsored workshop in the United States titled \textit{Research Assessment: What's Next?} and a 2000 workshop in Germany.

Unfortunately, ResEval “has not been very active” (G. Jordan, personal communication, March 7, 2006). WREN began with workshops which focused on the Office of Management and Budget’s (OMB) Research & Development Investment Criteria and Program Assessment Rating Tool (PART).\(^{43}\) In Korea, a World Research Evaluation Network was chartered in 2005, but nothing has yet emerged from it (G. Jordan, personal communication, March 7, 2006).\(^{44}\) All of these organizations have been focused on the evaluation of publicly-funded research and development (R&D),\(^{45}\) and in the last two decades large-scale research assessment systems have been developed and deployed almost worldwide (ab Iorwerth, 2005; Coryn, 2006, October, November; DEST, 2005; Campbell & Felderer, 1997; Geuna & Martin, 2003; RAE 2008, 2005; Tertiary Education Commission, 2005; University Grants Committee, 2004; VSNU, 2003).\(^{46}\)

\(^{43}\) PART was developed to assess the effectiveness of federal programs and to help inform management actions, budget requests, and legislative proposals directed at achieving results, and the tool’s developers argued that randomized controlled trials (RCTs) are the best type of evaluation design to demonstrate program impact (U.S. Office of Management and Budget, 2006a).
\(^{44}\) Recently, Japan started the process of developing the Japan Research Evaluation Network (JREN) (O. Nakamura, personal communication, December 16, 2006).
\(^{45}\) Recently, efforts at theory-driven and systems approaches for evaluating publicly-funded research have started to surface (Arnold, 2004; Molas-Gallart & Davies, 2006), as well as scientific and human capital models (Bozeman & Dietz, 1999, September).
\(^{46}\) A sample of these large scale systems used to evaluate government-funded research are discussed in Chapter II.
The Continuum of Complexity and the Governance of Research

As previously noted, evaluation has been an essential element of the scientific enterprise since before the appearance of the first scientific journals; usually in the form of peer review. Indeed, researchers themselves claim that they are the most frequently evaluated professional group. In the past few decades, however, the evaluation of scientific research, and in particular researcher performance, has changed substantially in terms of both scale and scope (Frederiksen, Hansson, & Wenneberg, 2003), as well as methodology. In part, these changes have occurred as a result of attempts to steer, regulate, and control research agendas and priorities, not only in regards to distributing research funds, but also to influence what occurs in the scientific system itself (Hansson, 2006). If there is a single word to describe this growth and change it is 'governance.' Hansson (2006) describes this type of research governance as "a somewhat ambiguous term for social regulatory processes that directly or indirectly implicate the political system; it is analogous to the sociologists' term 'social control'" (p. 174). Its political sweep is captured in the well-known admonition that what we need is "more governance and less government."

These transformations have resulted in what Lars Frederiksen, Finn Hansson, and Søren Wenneberg (2003) refer to as a 'continuum of complexity.' That is, research evaluation has evolved from the traditional and relatively simple peer review system to highly sophisticated benchmarking evaluation involving ever-growing numbers of criteria and standards, as shown in Figure 1. Thus, as a function of time, notions of research quality have become increasingly sophisticated. Research is no longer evaluated along the former one-dimensional criterion of its contribution to knowledge, but also along criteria evolving from increasing demands for responsibility and accountability, for example.
In part, these changes can be attributed to a radical departure from Vannevar Bush’s (1946/1960) notion of a linear model of economic and scientific growth. Compared to the linear model and as part of reforms in the public sector and within the concept of ‘the audit society’ (Power, 1997), the new complexity in research evaluation has emerged because:

...society is no longer trusts a one-dimensional evaluation approach conducted in an isolated system with one-dimensional and intrascientific criteria. The increase in research evaluations, the many new forms, and the many new criteria can be viewed as an attempt to create public trust in research—an attempt that results in a significant increase in complexity (Frederiksen, Hannson, & Wennberg, 2003, p. 156).

---

47 The linear model proposed by Bush (1960) asserted that the most efficient way to organize investment in science was to influence the distribution of money but to leave what goes on in science to the scientists.
Therefore, criteria other than strictly scientific criteria (e.g., social and political criteria) have been introduced, which further increase complexity, due to the new stakeholders of research evaluation who include: governments; politicians; the media; social movements; non-governmental organizations (NGOs); business people; and lay citizens (to include everyone). In addition to requiring new criteria, the changes have included the introduction of massive quantitative systems for “counting almost everything” (Frederiksen, Hannson, & Wennberg, 2003, p. 155), described as the ‘trust in numbers’ phenomenon (Porter, 1995). Beyond scale and scope (and methodology), these changes have also included new functions for evaluating research, such as the control of expenditures (i.e., value for the money) and in the decision-making context to legitimize decisions (Albæk, 1996; Hansson, 1998), among others.48

Retrospective

As evident in this brief overview, the systematic explicit evaluation of research has been a slow-developing endeavor. In the same manner, it was a long run from ancient science to sophistication about scientific method, and from there to the beginnings of serious assessment of the quality of the work of scientists, groups of scientists, and movements in science. However, the major claim from the above review is that for the entire scientific endeavor, science has value only because science is an evaluative endeavor itself. Indeed, since the institutionalization of science in the late 18th and early 19th centuries, evaluation has been central to research (Guena & Martin, 2003). That said, there are still serious weaknesses in many applications of the logic of evaluation to research; whether for assessing a researcher’s performance, for formulating research strategy, or for funding allocations to research, for example. The next section of this

48 See the Purposes of Research Evaluation section of this chapter.
chapter is devoted to that logic—which is intended to demonstrate the reasoning on which this dissertation is founded and with which it positions itself—and followed by the primary reasons and motives for evaluating research.

Logic of Evaluation

The logic of a subject or discipline concerns such matters as its definition and the definition of its major concepts, the nature of its relations to other subjects and other disciplines, and the rules of inference that govern it. No single scholar has contributed more to the logic of evaluation than Michael Scriven (e.g., 1969, 1982, 1991, 1993, 1995, 1996, 1997, 1998, 1999a, 1999b, 2005a, 2005b, 2005c, 2005d, 2005e, 2005, October, 2006a, 2006b, 2006c, 2006d, 2006e, 2006f, 2006g, 2007a, 2007b) and much of the following is drawn from his extensive work on the subject, in particular his *Key Evaluation Checklist* (KEC) which was originally developed in the early 1970s for use by Educational Testing Services (ETS) to evaluate education products produced by federally-funded R&D centers in the United States (Coryn, 2006b), and his *Evaluation Thesaurus* (Scriven, 1991).

Definitions of Evaluation

The common meaning of *evaluation*, as produced through a synthesis of dictionary definitions, of professional, disciplined evaluation yields the following:

49 Parts of this section appeared in the *Journal of MultiDisciplinary Evaluation* (Coryn & Hattie, 2006) and were presented to the Japanese Evaluation Society (Sasaki & Coryn, 2006, December).
50 During the writing of this section Scriven revised the KEC (2006g); however, most of the work contained herein was developed on the basis of the 2005 version (Scriven, 2005c).
51 In 1977, the KEC was more widely disseminated via Scriven’s *Evaluation Thesaurus* pamphlet, and subsequently through four book editions (Coryn, 2006b). Currently, Scriven’s revised versions of the KEC are published and distributed through The Evaluation Center’s (EC) Checklist Project which can be found at http://www.wmich.edu/evalctr/checklists/.
systematically and objectively52 “determining the merit, worth, or significance of things” (Scriven, 2005d, p. 235).53 Merit, worth, and significance are referred to here as the basic evaluative predicates.54 Broadly, merit (or quality) can be understood as intrinsic excellence, whereas worth (or value) is the extent to which the evaluand or valuee provides merit or quality under consideration of context and costs, and significance (or importance) is the overall conclusion when all relevant considerations have been synthesized. Determining merit, or worth, or significance is normally a process of measurement, observation, judgment, and/or inference. In addition, this definition implies that evaluators make evaluative claims, or conclusions, beyond the typical empirical or research-based claims of “what’s so?”55 to inferences or evaluative claims of “so what?”56 (Davidson, 2005b, p. xi).

Evaluative claims of the “so what?” variety are couched in value-imbued language such as good or bad, priceless or worthless, and trivial or important, which are subsumed in the vocabularies of merit, worth, and significance. In the present context, the evaluation of researchers and their research, it may be necessary to add additional elements, namely the “now what?” “how much?” “what if?” and “to whom?” as much of the evaluation of research involves questions related to the apportioning of research funding, setting of national research priorities and agendas, increasing or decreasing the scale of research initiatives, or predicting future performance, for instance.57

52 Meaning “as free from bias as we can make it within the budgetary and time constraints of the case” (Scriven, 2006c, p. 1).
53 As will be seen in the Definitions of Research subsection of this chapter, the definition of something is not a trivial matter and plays a crucial role in not only characterizing the essential nature of something, but also in the ideologies associated with the concepts on which the definition is founded.
54 Or, the “the triumvirate values of evaluation” (Scriven, 2006g).
55 Understanding the differences and connections between evaluation and other kinds of research and investigation, especially description, classification and diagnosis, generalization, prediction, explanation, justification, and recommendation is crucial to understanding the logic of evaluation.
56 For many years “so what?” questions were discounted on the premise that there were no scientific or rational answers and that evaluative questions were simply beyond the domain of scientific practice and reason (Scriven, 1991, 2005d).
57 The “now what?” “how much?” and “what if?” aspects necessary for some types of research evaluation often require some form of prediction. However, prediction is not the usual function of evaluation (i.e.,
Collectively, questions of this type are referred to as prospective (GAO, 1990), and they can be distinguished from questions about what is happening now or what has happened in the past; that is, retrospective questions.\textsuperscript{58}

To be useful a definition should be neither too narrow nor too broad, and like many disciplines evaluation suffers from both. For instance, authors of many of the most widely used evaluation textbooks define evaluation too narrowly as applied social science, such as in Rossi, Freeman, and Lipsey's (1999) \textit{Evaluation: A Systematic Approach}, where evaluation is defined as “the use of social science research procedures to systematically investigate the effectiveness of social intervention programs” (p. 4).

By contrast, the Organisation for Economic Co-Operation and Development (OECD) designates evaluation as:\textsuperscript{59}

The systematic and objective assessment of an on-going or completed project, program or policy, its design, implementation and results. The aim is to determine the relevance and fulfillment of objectives, development efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors.

Evaluation also refers to the process of determining the worth or significance of an activity, policy or program. An assessment, as systematic and objective as possible, of a planned, on-going, or completed development intervention (OECD, 2002a, pp. 21-22).

determining the merit, worth, and/or significance of something) and usually requires additional operations or tasks.

\textsuperscript{58} According to the United States General Accounting Office (GAO) Program Evaluation and Methodology Division (PEMD) prospective evaluation (i.e., forward-looking), can be described as a method for “providing the best possible information on, among other things, the likely outcomes of proposed programs, proposed legislation, the adequacy of proposed regulations, or top-priority problems” (GAO, 1990, p. 1).

\textsuperscript{59} The OECD definition also excludes personnel and product evaluation (see the Fields of Evaluation subsection of this chapter), as well as costs and generalizability (see the Core Dimensions of Evaluation subsection of this chapter).
While the OECD defines evaluation in terms of particular uses (e.g., assessment of effectiveness, impact, and sustainability), Trochim (2001), on the other hand, broadly defines evaluation as ‘what evaluators do.’

Many types of evaluations do not necessarily result in an assessment of worth or merit—descriptive studies, implementation analyses, and formative evaluations, to name a few. Better perhaps is a definition that emphasizes the information-processing and feedback functions of evaluation...one might say that evaluation is the systematic acquisition and assessment of information to provide useful feedback about some object (Trochim, 2001, p. 30).60

By contrast, proponents of theory-driven (Chen, 1990, 2005a, 2005b) and realistic (Pawson & Tilley, 1999) evaluation approaches characterize evaluation as explaining how and why programs work, for whom, and under what conditions, giving almost no mention to merit, worth, or significance (Coryn, 2005, 2006c), while advocates of the empowerment evaluation movement portray evaluation as “the use of evaluation concepts and techniques to foster self-determination” (Fetterman, 1994, p. 1).

A myriad of definitions and conceptions of evaluation have been put forward in handbooks, guidelines, and administrative procedures (Stern, 2005), of which the aforementioned (Chen, 1990, 2005a, 2005b; Fetterman, 1994; OECD, 2002a; Pawson & Tilley, 1999; Rossi, Freeman, & Lipsey, 1999; Trochim, 2001) are only a few of the many instances where the essential nature of evaluation has been definitionally misconceived. Of course, part of the problem is that most definitions define evaluation only in terms of program evaluation. The result of these conflicting definitions has been that:

...standardizing on their own (significantly different) usage is of course just the kind of confusion at the macro level that the standardizers are trying to avoid in their own bailiwick: a person learning or using one set of definitions will have trouble understanding and communicating with those trained to another version (Scriven, 2004, p. 13).

60 Trochim’s (2001) definition would include almost all forms of descriptive research.
In any case, the amalgamation of dictionary definitions should not be abandoned and going beyond *systematic and objective determination of merit, worth, or significance* evaluation is essentially nothing more than claims of personal preference.

Evaluative Claims

Evaluative claims are what distinguish evaluation from the usual research paradigm (Gugiu, 2006, November); that is, getting from the “what’s so?” to the “so what?” There are at least four varieties, or types of evaluative claims, which are (Scriven, 2005d).\(^{61, 62}\)

*Personal preference claims.* Type I value claims. Personal preference types of evaluative claims are one of the sources for the fallacious belief that science is, or should be, value free. These types of claims are of the “I like” or “I favor” variety and express values that are not definitionally true and cannot be validated as having any wider applicability. Normally, these types of claims are connected to a private state of the mind; that is, *subjective*.\(^{63}\)

*Market value claims.* Type II value claims. Market types of evaluative claims are those that have a standard method of verification recognized by the law and common sense; for example, the price that an interested, but not desperate, buyer would be willing to pay and an interested, but not desperate, seller would be willing to accept on the open market. These types of values are testable and much less subjective than the personal preference variety given that they have no direct connection to a personal state of the mind.

*Contextual value claims.* Type III value claims. Contextual types of value claims are those that are *prima facie factual* but, normally, only in a certain context that refers to properties that are highly valued in that context and hence carry the import of an inference of value. In the realm of the

---

\(^{61}\) In earlier works Scriven (1991) argued that there were essentially two types of evaluative claims: *primary* and *secondary*. Primary evaluative claims were of the sort that something had a certain merit or value, whereas secondary evaluative claims were of the sort that someone, including oneself, believed that something had value. These were eventually replaced with the more fully developed types presented here.\(^{62}\) Schwandt (2005), however, asserts that the four types of value claims are aesthetic, moral, utility, and instrumental.

\(^{63}\) These are the types of claims, in general, which serve as the basis for relativist and other assertions that there is no objective truth, or more narrowly, that there is no basis for evaluative claims of merit or worth.
evaluation of research ‘prolific’ could be considered a contextually evaluative claim, or that ‘internationally-relevant’ research has greater value than ‘locally- or nationally-relevant’ research could be considered a contextually evaluative claim, or that only research conducted ‘within-paradigm’ has value could also be considered as a contextually evaluative.

Essentially evaluative claims. Type IV value claims. Essentially evaluative types of value claims are those which professional evaluators strive for and aim to support and are entirely unlike those of the personal preference variety. These types of claims “in some cases lack the support that entitles them to be called provable, in many other cases, they have more than enough support to justify the view that they are as well supported as the usual kind of particular or general scientific claim” (Scriven, 2005e, p. 237). There is more than ample, objective and scientific, evidence to support and justify evaluative claims, of the essential variety, about the brilliance of Einstein, for example.

In order to arrive at evaluative conclusions, especially Type IV, it is usually necessary to establish or identify two kinds of premises, which are (Scriven, 1999a, 1999b):64, 65

Factual premises. Factual premises, in evaluation, are about the nature, performance, or impact of an evaluand or evaluee.66 These premises are roughly equivalent to description as it is understood in the empirical sciences; that is, “what’s so?”67 Good scientific description, and factual premises in evaluation, involves applying context-dependent standards of thoroughness and precision.68

Value premises. Value premises, in evaluation, are about the relevant values. These values serve as the basis for what is brought to bear, in combination with factual premises, to determine the merit, worth, and/or

64 In part, the distinction between facts and values is the distinction between the vocabulary of description and that of evaluation.
65 These two premises closely approximate Stake’s (1967) notion of the two countenances of evaluation, which are description and judgment.
66 The term evaluand means that which is being evaluated (e.g., a program, policy, product, portfolio, proposal). In personnel evaluation the term is evaluee.
67 In evaluation, description is often referenced to in terms of performance.
68 Factual premises, or description, in evaluation often involves the application of social and other scientific methods (e.g., measurement or observation), but also additional investigatig tools not normally thought of as social science or traditional scientific tools, such as needs assessment (for both factual and value premises), eliminative analysis, observational causal analysis, modus operandi methodology (Scriven, 1976), cost analysis, side-effect identification, and so on (Scriven, 2005a, 2006c).
significance of an evaluand or evaluee; that is, "so what?" Normally, value premises are the irrefutable properties or characteristics which typify a good evaluand or evaluee of a particular class or type in a particular context.

There are usually many, often dozens or several hundred, of these premises in the evaluation of complex entities or entities with complex functions (Scriven, 1999a). To obtain evaluative conclusions of the essential type it is typically necessary to combine or integrate all relevant factual and values premises by means of the synthesis operation (Davidson, 2005; Gugiu & Persaud, 2006, April; Scriven, 1991, 1994b, 2005c; Scriven & Davidson, 2000, November).

The Tripartite Taxonomy

There are normally three primary purposes, herein referred to as the tripartite taxonomy, for making evaluative claims or for conducting evaluation, which are: formative, summative, and ascriptive. Formative evaluation is typically conducted "during the development...of a program or product (or person, and so on) and is conducted, often more than once, with the intent to improve" (Scriven, 1991, pp. 168-169; Scriven, 1996). By contrast, summative evaluation "of a program (or other evaluand) is conducted after completion [or near the end] of the program...[and]...for the benefit of some external audience or decision-maker" (Scriven, 1991, p. 340; Scriven, 1996). However, formative evaluation can also be undertaken with some external audience in mind. Unlike

---

69 While factual premises describe performance, value premises are the standards applied to determine how good or bad, worthwhile or worthless, or significant or insignificant the evaluand's or evaluee's performance is or was.

70 Sources of relevant values are described in greater detail in the Values and Valuing subsection of this chapter.

71 Making evaluative claims, or arriving at evaluative conclusions, does not always or necessarily imply making recommendations, explanations, or predictions (Scriven, 1994a).

72 This synthesis operation is one of the key logical processes in evaluation and is a long way from the simple deduction and statistical inference that are more common elements in scientific inference (see the Fundamental Operations subsection of this chapter).
formative and summative evaluation, ascriptive evaluation is neither aimed at improvement nor at decision making, specifically, and is normally done merely for the sake of knowing; that is, ascriptive evaluation is roughly equivalent to Michael Q. Patton's (1997) and Chelimsky's (1997b) notion of evaluation's function to generate knowledge.\(^{73}\)

The formative and summative roles of evaluation are not always mutually exclusive, are occasionally orthogonal, and have been the subject of much controversy (e.g., Chen, 1996a, 1996b; Patton, 1996; Scriven, 1996). Nevertheless, one thing is clear, the logic and lexicography of evaluation “does require that both formative and summative evaluation involve efforts to determine merit” (Scriven, 1996, p. 157) and that this distinction is ultimately context dependent. An editorial decision to ‘accept’ or ‘reject’ a research manuscript submitted for publication is summative, while a decision of ‘revise and resubmit’ is formative. For the author, however, both the reject and revise and resubmit decisions can be formative in that the author can opt to improve the manuscript, especially if feedback was given by reviewers or the editors. In any case, the decisions are de facto summative in the editorial context, but nearly always formative in the context of the author. Of course, the author could simply make a decision to submit the manuscript to another journal, in which case the author has undertaken a summative evaluation of another kind—a decision not to make use of the editors’ recommendations (i.e., revise and resubmit) as a basis for improvement.

\(^{73}\) Ascriptive evaluation first appeared in Scriven’s (2004) editorial, *The Fiefdom Problem*, in the 1st issue of the *Journal of MultiDisciplinary Evaluation*. Ascriptive evaluation is “an evaluation done simply in order to determine the merit, worth, or significance of the evaluand. Contrasted with formative and summative evaluation, which are done to assist, respectively, program developers and program ‘disposers’ (i.e., decision-makers about the fate or funding of the program). Sometimes the term ‘knowledge-oriented’ is used to identify what are here called ascriptive evaluations, but that’s a false contrast, since the results of formative and summative evaluations are also knowledge. What is sometimes called diagnostic evaluation is a sub-species of ascriptive evaluation, since the diagnostic categories used in evaluation (e.g., clinical disorders in psychotherapeutic evaluation, learning disorders in educational psychology) are all evaluative categories, so one is determining the particular variety of evaluative state of the evaluand” (M. Scriven, personal communication, February 21, 2006).
Values and Valuing

The “so what?” type of evaluative claim “brings us face to face with the most central question in the logic of evaluation...how is it possible to justify answers to questions about value in a scientific or other disciplinary way?” (Scriven, 2005d, p. 236).74 In evaluation, value definitionally refers to something, which is in principle or quality, intrinsically valuable or desirable.75 The most serious threat to ‘valuing’ in evaluation, is that in making evaluative claims one is committing the ‘naturalistic fallacy’,76 which is often purported to establish the impossibility of objective scientific demonstrations of evaluative conclusions based on values.77 This shadow has been predominately cast by the pervasive remnants of Weber’s ‘mantle of objectivity’ in the social sciences and the pernicious effects of the 20th century positivist philosophy, among others, which assert that:

...value claims are (merely) expressions of feelings or attitudes of approval or assertions of will. They exemplify moral subjectivism, the belief that moral positions are not grounded in reason or in the nature of things; rather, we simply adopt them because we are drawn to them emotionally. Ultimately, value claims are matters of choice not grounded in rationality, and, as such, are outside the realm of scientific investigation (House & Howe, 1999, p. 5).

Although many would argue—or assume—that the value-free doctrine has been abandoned, this is simply not the case:

74 By contrast, other evaluation approaches, such as Eisner's model of connoisseurship and criticism (Donmoyer, 2005; Eisner, 1976) and Guba and Lincoln's Fourth Generation Evaluation (Guba & Lincoln, 1989; Lincoln, 2005) describe valuing not as an objective or scientific procedure, but rather as personal preference, expertise, or social constructions.
75 House and Howe (1999) define value as “a concrete noun to refer to what has value or is thought to be good—for example, democratic values, conservative values, or stakeholder values [italics in original]” (p. 6).
76 The naturalistic fallacy, an alleged logical fallacy, was introduced by the British philosopher George Moore in Principia Ethica (1903/1993), in which it was stated that a naturalistic fallacy was committed whenever a philosopher attempts to prove a claim about ethics by appealing to a definition of the term 'good' in terms of one or more natural properties such as 'pleasant,' 'healthy,' 'natural,' and so forth.
77 For a complete discussion of the fallacies of the value-free doctrine see Scriven (1991, 1993).
...[this] view that we have outgrown the values-free doctrine is dangerous because it is based on a complete misrepresentation of the doctrine, and hence on pseudorefutation. The doctrine has not in fact been discarded by many, probably most, of the methodologically sophisticated social scientists who are leaders in their fields—certainly not in their private thinking (Scriven, 1993, p. 12).

If true, the value-free doctrine is of essential importance to evaluation, and the evaluation of researchers and their research, since "it would invalidate almost any claims to objective evaluation" (Scriven, 1991, p. 373). Nonetheless, despite the claims that 'science is only descriptive' and 'values are always subjective,' scientists and researchers themselves are notorious for making evaluative claims; for instance, about the merits of prior contributions to the scientific literature, of explanations, of the predictive power of theories, of fit, of the quality of data, of research designs, and of interpretations of research results. These types of evaluative claims demonstrate that the value-free doctrine is self-refuting given that the scientists and researchers who make them do so on the basis that these conclusions are supported and justified by factual, value, and definitional premises, and more often than not "the value premises turn out to be not arbitrary expressions of taste but instead perfectly acceptable and actively defended positions" (Scriven, 1993, p. 13), and therefore relatively uncontroversial.

It is clear then, that values can be established scientifically and objectively, in principle and practice, once it is realized that a very substantial portion of the sources of values, as applied in evaluation, are drawn from demonstrably defensible sources (Youker, 2006, November). However, the identification of relevant value premises requires

---

78 It is also essential to understand that Weber's value-free doctrine has nothing to do with the "straw-man version" (Scriven, 1991, p. 373); the claim that scientists' activities and their conclusions should never be affected by their personal or cultural values.

79 The term 'value' literally means 'the judgment of what is important.'
great care,\textsuperscript{80} and in most cases come from one or more of the following sources (Scriven, 2005c, 2006g):

\begin{enumerate}
  \item \textit{Needs assessment.}\textsuperscript{81} Needs assessment, where need is usually defined as anything essential for a satisfactory mode of existence or level of performance, is a diagnostic process for discovering the facts about the functions or dysfunctions of organisms or systems.\textsuperscript{82} Needs are often the root source for values by which an evaluand's or evaluatee's performance is upheld.\textsuperscript{83} In evaluating researchers and their research, this source of values, or criteria, is not always a straightforward matter, and may require some distinction between basic and applied research. That is, on the one hand, basic research may not serve an immediately observable need, although it may contribute to a future need in the form of knowledge, for example. On the other hand, applied research may address a presently unmet need or attempt to improve the way in which already met needs are addressed.\textsuperscript{84}
  \item \textit{Definitional.} Definitional values, or criteria, are normally drawn from the evaluand or evaluatee, or from standard usage (Caryn, 2006d). These values often relate to, for instance, breadth or depth of impact.\textsuperscript{85} As applied to research, this might consist of values related to economic or social benefits produced as a result of a research product or the contribution to knowledge (e.g., a description or explanation of part of the world).\textsuperscript{86}
  \item \textit{Logical.} Logical values are those that are formally true or valid, in accordance with principles of reasoning on the basis of inference and demonstration. Logically, values which define the properties
\end{enumerate}

\textsuperscript{80} The demonstrably relevant values to be applied in the case of most researchers and their research are enumerated in the \textit{Properties of Good Research} section of this chapter.

\textsuperscript{81} Evaluations build on needs assessment as "theories build on observation," that is, "it's not that observations are infallible, only that they're \textit{less} fallible than theoretical speculation" (Scriven, 1991, p. 243).

\textsuperscript{82} Conceptually, needs can be distinguished as conscious or unconscious, met or unmet, and performance or treatment types of need, as well as from wants or desires (Davidson, 2005b; Scriven, 1991).

\textsuperscript{83} Assessing needs often requires a long-term outlook and needs assessments are normally conducted within a contextual framework (Altschuld & Kramer, 2005; Altschuld & Witkin, 2000; Coryn, Gugiu, Davidson, & Schröter, 2007; Davidson 2005b; McKillip 1998). Once identified, needs may be used to develop explicit criteria by which evaluators can determine how well a program is meeting those needs.

\textsuperscript{84} For example, applied research on tissue regeneration or renewable energy sources are usually self-evident needs in some, but not all, cases, whereas basic research in music on the properties of harmonic discordance may never meet an observable need.

\textsuperscript{85} For example, "a program is usually regarded as definitionally better if it reaches more people and has a larger good effect on them" (Scriven, 2005c, p. 4).

\textsuperscript{86} See the \textit{Definitions of Research} subsection of this chapter.
of good research might include, among others, innovativeness, profundity, creativity, or originality.

iv. **Legal.** Legal values, or requirements, are normally those that are legislated or mandated. For instance, in the case of research, federal regulations, state law, and sponsor requirements for the use of human subjects in research.

v. **Ethical.** Ethical values, or requirements, are those that relate to relevant ethical requirements, not covered under legal values. For researchers or research this might consist of not publishing plagiarized work, falsifying findings, or causing harm to research subjects.  

vi. **Personal and organizational goals/desires.** Personal and organizational goals and desires are values which are normally much less important than the needs of consumers, since they often lack ethical justification. However, in the research case this might cover matters related to esteem or scientific prestige of a nation’s research, a research institution, a research group, or an individual researcher. In goal-free evaluation, personal and organizational goals and desires are normally irrelevant (Evers, 1980; Schröter, Coryn, & Beywl, 2006, October; Scriven, 1972, 1973, 1991; Youker, 2005).

vii. **Fidelity.** Fidelity, or authenticity, adherence, or compliance values are those normally related to implementation. For example, the extent to which funded research was executed as agreed upon or regard for the assumptions of a specific research design; for example, serious threats to validity in randomized controlled trials.

viii. **Sublegal.** Sublegal values are those not normally covered by legal or ethical values, in reference to important, often idiosyncratic, legislative preferences as opposed to those which are mandated. In the evaluation of researchers and their research, this might include for example, safeguards for keeping research data confidential or destroying data after a specified period of time has elapsed.

ix. **Professional standards.** Professional standards are the values established by authority, custom, or general consent; most often

---

87 Value sources iv (Legal) and v (Ethical) often, but not always, overlap.

88 Furthermore, if the goals are not worth achieving then it is unimportant to determine how well they were met (Scriven, 1967).
as determined by a profession or professional association (Frankel, 2004). For researchers these standards are frequently drawn from field- or subject-specific standards of conduct or practice. 

x. **Expert judgment.** Expert judgment values, usually those of subject matter experts, normally refer to the preferred standards of "experienced practitioners" (Picciotto, 2005, p. 31). However, expert judgment often takes the form of connoisseurship and is also subject to the fallacy of irrelevant expertise, among others. 

xi. **Historical/traditional/cultural standards.** Values drawn from historical, traditional, and cultural standards are normally time and context dependent. For research, these values might include within-paradigm or methodological traditions (i.e., historical or traditional standards) or the values associated with Māori and Pasifika research in New Zealand (i.e., cultural standards), for example. 

xii. **Scientific.** Scientific merit, worth, or significance usually refers to the contribution to knowledge made by a researcher, research group, or institution, for example, and/or the research product thereof. In the entire research endeavor this is normally considered one of the single most important values, or criteria, for determining whether a researcher or their research is meritorious, valuable, or important. Less frequently, this value may refer to methodological rigor or fitness for purpose, for example. 

xiii. **Technological.** Technological merit, worth, or significance, refers to a process and also product of that process. In the case of technology, the product is the artifact and the process is whatever it takes to make the product. Numerous values or criteria may

---

89 For example, a professional standard may be related to conduct (however, this normally falls under legal or ethical dimensions), licensure, or certification.

90 The American Psychological Association's (APA) *Ethical Principles for Psychologists and Code of Conduct* (2003), for example, has professional codes of conduct not only for practicing psychologists, but also psychological scientists and researchers, which includes codes of conduct in the care and use of animals for research purposes.

91 In more traditional forms of program evaluation this value source carries a somewhat different meaning. For example, the program design is based on sound scientific evidence.

92 For instance, in R&D evaluation, the process of identifying and validating values, or criteria of technological merit, worth, or significance for a product of R&D originates with criteria derived from needs assessment.

93 R&D efforts, aimed at breakthroughs, require: (i) serious background research; (ii) planning that covers all worst case scenarios; (iii) many field testing cycles; (iv) tough evaluation; (v) multiple solutions; (vi) scaling; and (vii) cost analysis (Scriven, 1991).
emerge here, including social change, for example. Often, this value is aimed at total impact. In the case of researchers or their research, this normally applies to R&D evaluation or technology assessment.

xiv. **Marketability.** Marketability values are those that refer to potential user's access to the product, or program, or the plan for getting them used. For instance, research may be *prima facie* important, but if it does not reach those who are in need, those who would benefit from it, or those who would use it, then no needs are met, no benefits are produced, nor is it utilized.

xv. **Political.** Political values are those—sometimes—narrowly or widely accepted dimensions, which must—sometimes—be considered relevant, without the pejorative connotations normally associated with these types of values. Political values are extremely tricky, and for instance, in the case of research, might include values such as nation's position in research productivity relative to other nations.

xvi. **Resource economy.** Resource economy values are those that are associated with economic or social impacts or benefits with respect to monetary and non-monetary costs, or similar results which could have been produced with lesser resources, for example. Resource economy is a crucial value in the evaluation of proposed research in particular; for example, can the same, or similar, results be obtained at lower costs?

xvii. **Risk.** Risk or risk-aversion values, on the one hand, subsume costs in the best, worst, and most likely cases, not just the latter, in order to give a full and true account of the costs. On the other hand, risk is also the potential harm that may arise from some present process or from some future event. In everyday usage risk is often used synonymously with probability, but normally risk combines the probability of a negative event occurring with how harmful that event would be. Risk is also associated with uncertainty (Tsipouri, 2006, November).

---

94 Resource economy often includes opportunity costs as well (see the Costs Evaluation and Comparative Evaluation subsections of this chapter).

95 In scenario analysis risk is distinct from threat. A threat is a very low-probability but serious event—which some analysts may be unable to assign a probability in a risk assessment because it has never occurred and for which there is no available preventive measure. The difference is most clearly illustrated by the *precautionary principle* which seeks to reduce threat by requiring it to be reduced to a set of well-defined risks before an action, project, program, innovation, or experiment is allowed to proceed (Holton, 2004).
It is also important to distinguish between two types or classes of values as applied in evaluation: *general values* and *specific values*. The former are the merit-defining *criteria* by which an evaluand or valuee is evaluated; the properties or characteristics which define a 'good' evaluand or 'good' valuee. The latter are the *standards* which are applied and by which performance is upheld, in order to determine if that performance is or is not meritorious, valuable, or significant.

In the evaluation of researchers, for instance, productivity is often considered a general value; that is, productivity is a dimension or property which defines a good or valuable researcher, though this is only one of many such properties. If productivity in this case were taken simply as the number of publications in refereed scholarly journals in the past two years, for example, and the researcher in this case had four publications in the previous two years, then converting the researcher's observed productivity to a value claim on the productivity dimension or criterion (i.e., whether the researcher's performance on the productivity criterion is poor, average, or excellent, for example), requires a standard—the specific value—by which that performance is upheld and the standard is brought to bear.

If the researcher happened to be a cognitive psychologist, and the average number of publications in that disciplinary subfield is two per year, or four in two years, then the researcher in question could be classified as 'average' on the productivity criterion. If that researcher had seven publications the previous two years, as opposed to four, then the researcher might be considered 'excellent' or 'outstanding' on the criterion.

---

96 Criterion literally means 'a standard on which a judgment or decision may be based' or 'a characterizing mark or trait.'

97 These are not stylistic standards, which are predicated on voluntary compliance (Picciotto, 2005). For example, the Joint Committee's standards for program and personnel evaluation are those by which evaluators should—voluntarily versus mandatorily or obligatorily—adhere (1988, 1994).

98 See the *Properties of Good Research* section of this chapter for additional criteria which define good research.

99 This example is an intentional oversimplification and discounts journal impact factors and other relevant standards related to scientific productivity.
of productivity, which in part depends upon the rubric\footnote{Rubrics are commonly employed as tools for converting factual, observed performance to grades, scores, or ranks in evaluation (Davidson, 2005b; Stevens & Levi, 2005). In this case, a simple rubric for converting observed performance (i.e., number of publications in the previous two years) to a merit rating could be: 0 publications = unacceptable (or F); 1-2 publications = poor (or D); 3-4 publications = average (or C); 5-6 publications = good (or B); and > 6 publications = excellent (or A).} or other method employed to convert observed performance to standards of merit. But, as will be seen later, the quality of these four or seven may play an important role in modifying this judgment.

Criteria versus Indicators

Given that the path of righteousness for evaluators is the path of criteria, not indicators, how do we identify true criteria for evaluand X?...what properties are part of the concept of a “good X”? (Scriven, 2005e, p.57).

Much of the confusion in evaluation, as it has been historically and as it is currently practiced, stems from the failure to adequately distinguish between criteria and indicators.\footnote{Furthermore, Stake, Migotsky, Davis, Cisneros, DePaul, Dunbar, Farmer, Felto, Johnson, Williams, and Chaves (1997) contend that procedures for recognizing quality directly via criteria are a long way from common practice.} \footnote{It is also useful to distinguish between standards and criteria (Glass, 1977).} Indicators are factors, variables, or observations that are empirically connected with a criterion variable; for example, a correlate. Criteria, by contrast, are definitionally connected with the evaluand (Scriven, 1959). That is, criteria are the properties or characteristics that delineate a good, valuable, or significant evaluand or evaluatee of a particular class or type.\footnote{See the previous subsection, \textit{Values and Valuing}, for sources of relevant values or criteria.} Indicators, unlike criteria are frequently, but not always, unstable in their validity, and many cases they are “easily manipulated” (Scriven, 1991, p. 194).\footnote{Some indicators, however, are very stable and difficult to manipulate. For instance, breathe analysis as an indicator of poor driving. Being over the legal blood alcohol concentration (BAC) limit (usually $\geq 0.05$) is nothing but a correlate and certainly not a criterion, but is very stable and difficult to manipulate.} Most credible and valid merit-defining criteria are developed on the basis of the following seven requirements:

1. \textit{Critical status}—not merely indicators
2. Completeness—no significant omissions
3. Non-overlapping—discreteness and independence
4. Commensurability—equitable levels of generality
5. Clarity—comprehensibility and applicability
6. Conciseness—brief in statement or expression
7. Conformability—measurable or reliably observable

It is also useful to unpack merit-defining criteria into subcriteria (or subdimensions), or in some cases even further (Coryn, Hanssen, Gullickson, & Ritchie, 2005, October; Schröter & Coryn, 2005, November). This process often adds richness, depth, and value to a list of criteria of merit, particularly when used for formative evaluation. For example, in evaluating a proposal to conduct research, or a manuscript submitted for publication, it is more useful to the evaluand (or evaluatee) to know that the intellectual merits were rated high and that the significance aspect was rated low, rather than simply ‘rejecting’ the manuscript. Without such information it is much more difficult for the evaluand to make improvements to current or future work.

Fundamental Operations

There are essentially five fundamental or core evaluative operations, with some minor variations, used for determining the absolute or relative merit, worth, or significance of an evaluand or evaluatee. The basic operations employed in evaluation are grading, ranking, scoring, apportioning, and synthesis. In most cases, the evaluation of research or researchers involves grading and ranking. However, as will be seen in Chapter II, it also frequently involves ranking and apportioning.105

---

105 Grading and ranking are often used as a basis for apportioning of research funding.
Grading. Grading, as used in evaluation, is an operation that involves assigning evaluands to an ordered set of categories, with the order corresponding to a metric of merit. Most often categories are assigned a letter grade to represent merit, such as A = excellent, B = good, C = satisfactory, D = poor, and F = unacceptable, for example. In cases where the number of evaluands or evaluatees is relatively small, and performance has a high degree of variation, grading can also be used to generate a ranking via pairwise comparisons. Grading is frequently used in the evaluation of research to assign a researcher, their research, a research proposal, a research group, department, or institution, to a category of merit.

Ranking. Ranking is an operation used to place evaluands or evaluatees in an order, for example, of merit, on the basis of their relative performance on a measurement or observation. Ranking is frequently used in the evaluation of research, for example, to set national priorities or rank a nation’s research or researchers in comparison to world norms. Ranking cannot, however, provide grading without additional assumptions about the metric of merit. Ranking can also be of the full, partial, or gap variety; where full ranking does not allows ties, partial ranking does, and gap-ranking specifies some distance, that is, interval, between rankings. Gap ranking is also useful for apportioning, in some contexts, of research funding, for instance.

Scoring. Scoring involves assigning numeric quantities, usually in terms of performance, on which to represent merit. Points are usually supposed to be of equal value, awarded for meritorious performance, and may also be used to represent a numerical grade and in some cases used to rank within grades. Converting scores to grades requires a point constancy requirement. Scoring has, in the past, been employed in evaluating the quantity, or in some cases quality, of a researcher’s performance, on the basis of simple counts of publications where points are awarded on the basis of publication type (e.g., book review = ½ point, journal article = 1 point,

---

104 Using pluses (+) and minuses (-) in grading merely creates additional categories of merit.
105 For example, using a pairwise comparison procedure for ranking graded candidates, each candidate, or alternative candidate, is matched head-to-head, or one-on-one, with each of the other candidates. Each candidate and alternative candidate receives one point for a one-on-one win and a half a point for a tie. The candidate and alternative candidate with the most total points is the winner, that is, ranks first. A similar procedure has been used to rank order engineering designs based on categorical quality ratings (Dym, Wood, & Scott, 2002).
106 For example, one’s research could easily be ranked 1st among a group of researchers, but still not meet the criteria for an A grading.
107 Gap-ranking, using a ‘horse race’ analogy, is used to represent some distance, or estimated interval, between rankings such as ‘by a nose,’ ‘by a head,’ or ‘by three lengths,’ for example.
110 A point constancy requirement, when used for numerical scoring, is essentially that a point should reflect the same amount of merit, however earned.
Apportioning. Apportioning, also frequently referred to as allocation or distribution, is the process of dividing a given, often finite, quantity of valued resources between competing demands. Apportioning is a unique evaluative operation, distinct from grading, ranking, and scoring. Although, these (grading, ranking, and scoring) are all involved, apportioning may be "logically reducible to a very complex combination of grading and ranking" (Scriven, 1991, p. 58). The apportioning operation is one of the most important in the evaluation of research, particularly in cases of questions or purposes of the "now what?" and "how much?" variety. For example, "How much funding should be allocated to nuclear engineering? How much to genetic engineering? How much to theoretical physics?"

Synthesis. The synthesis operation is the process of amalgamating a set of ratings or performances on several dimensions, components, or criteria into an overall evaluative conclusion (Persaud, 2006, November; Scriven, 1994b; Scriven & Davidson, 2000, November) or the process of combining factual and value premises (Gugiu, 2006, November). It can also be the process of "combining a set of ratings or performances on several subdimensions into a rating on one dimension" (Davidson, 2005b, p. 248). Often, the synthesis operation is conducted using a numerical weight and sum (NWS) or qualitative weight and sum (QWS) methodology (Davidson, 2005b; Scriven, 1991). NWS usually involves applying an algorithm, whereas QWS usually involves applying a heuristic. The synthesis operation can be applied for determining either absolute or relative merit and/or worth. Synthesis is the inverse operation

---

111 Simple publication or citation counts are rarely used in contemporary research evaluation practice on the grounds that they (i.e., quantity of publications) are generally considered spurious measures of research quality.

112 Apportioning is often said to be the defining problem of the science of economics. However, it is rarely addressed in practical terms, in part, due to the interpersonal comparison of utility. That is, "does a dollar count as having the same value no matter to whom it is allocated?" (Scriven, 1991, p. 198).

113 In a growing number of countries, government-appointed assessment panels develop ranks on the basis of the quality of scholarly outputs to apportion budgets in recognition of evaluated performance and to justify expenditures of public funds (Coryn, 2006, October, 2006, November; Lange, 2006).

114 Apportioning is more evaluative than ranking because one is often inclined to say that it ought to be done evaluatively. "If the money allocated for merit raises is given out to the faculty in history simply proportionately to their current salary, that is, on a percentile basis, we are inclined to say that it is improper. But the normal raises across the board, often called cost of living raises, are apportionments of money and they are not quite improper, though questionable," (M. Scriven, personal communication, July 21, 2006) given that the cost of living is not proportional to salaries.
of *analysis*, which essentially separates a whole into its component parts.\textsuperscript{115}

NWS involves ascribing, usually interval or ratio, numerical weights and numerical performance scores on each dimension, or subdimension, multiplying weights by performance scores, and then summing the products. The resulting sum represents the *overall merit* of the evaluand or evaluatee. Although the NWS procedure is widely used and appears commonsensical, it is, however, subject to the fact that no numeric weights can compensate for a minimum performance on some dimension (see *Barring*), it assumes equivalence in scoring differences, very often they consists of non-linear distributions of value within a scale or across scales, and good performance on trivial dimensions can often ‘swamp’ poor performance on crucial dimensions, for example (Scriven & Davidson, 2000, November). NWS may also require pairwise comparisons in some cases.

QWS is a non-numerical procedure, using only non-numeric (e.g., grading) scales for weighting and performance, where performances on multiple criteria are summed to determine overall merit or worth. It is a ranking methodology, most often used for determining the *relative merit* of two or more evaluands or evaluatees.

QWS is preferable to NWS for most synthesis operations due to a number of logical problems associated with the NWS operation (Davidson, 2005b; Scriven, 1991; Scriven & Davidson, 2000, November).

**Secondary Operations**

In addition to the five fundamental or core evaluative operations there are a similar number of equally important secondary operations. These secondary operations are weighting, barring, stepping, scaling, and profiling.

\textsuperscript{115} The synthesis operation also requires an understanding of the difference between holistic evaluation and three types of analytic evaluation: dimensional; component; and theory-driven. *Holistic evaluation* is not, explicitly, analytic and attempts to determine the merit or worth of an evaluand or evaluatee without consideration of the separate components or dimensions of merit. *Dimensional evaluation* is a form of analytic evaluation in which merit or worth are determined by performance on multiple dimensions which pertain to the evaluand or evaluatee as a whole rather than by its separate components (i.e., the sum of its parts). *Component evaluation* is a form of analytic evaluation in which merit or worth is determined by an evaluand’s or evaluatee’s performance on each separate component, which is then synthesized into an overall evaluative conclusion. *Theory-driven evaluation* is a form of analytic evaluation which essentially attempts to explain the causal relationships between an evaluand’s components and to answer how and why an evaluand achieves a result (Chen, 2005a, 2005b; Coryn, 2005).
Weighting. The weighting operation involves assigning levels of importance to components or dimensions of an evaluand or evaluatee to indicate their relative or absolute importance.\textsuperscript{116} Weights can be numeric or non-numeric (e.g., NWS and QWS operations). Numeric weights can be on a scale of 1-3, 1-5, 1-10, and so on, whereas non-numeric weights can be defined in terms of categories such as 'critical,' 'important,' or 'desirable,' for example. In evaluating higher education faculty it is not uncommon to give greater weight to research (e.g., weight $= 3$, or critical) than to teaching (e.g., weight $= 2$, or important) and/or to service (e.g., weight $= 1$, or desirable).

Barring. Barring is an evaluative operation where minimum levels of performance are set, or required, on specific dimensions or components, performance below which cannot be compensated for by better performance on other dimensions (i.e., minima).\textsuperscript{117} Therefore, failure to 'clear' a bar means 'failure' of the evaluand or evaluatee. In the case of research a bar may be placed on the \textit{ethicality} dimension, for example, where publishing fictitious research results,\textsuperscript{118} failure to report harmful side effects, or plagiarism may result in \textit{holistic}\textsuperscript{119} or \textit{global}\textsuperscript{20} failure, no matter how good performance is on other relevant dimensions.

However, the barring operation can be taken further still with the notion of the \textit{high} and \textit{low bar}—or, 'double bars.' On a relevant and heavily weighted dimension, such as intellectual brilliance for example, there are rare cases where one is 'off the scale' good (e.g., 'the Einstein case')—the high bar. It is not important that the evaluatee published only three papers, but rather that they were of such significance that the evaluatee 'grabbed'

\textsuperscript{116} The weighting operation, and to some extent the barring operation, requires logical justification and validation. Equal weighting (i.e., \textit{pari passu}) should not be abandoned without overwhelming evidence; however, equal weighting on each criterion or dimension is often an oversimplification.

\textsuperscript{117} Other barring operations include Davidson's (2001, 2005b) \textit{hard} and \textit{soft hurdles}. Hard hurdles require an evaluand or evaluatee to meet an overall passing requirement, and "if the evaluand or evaluatee fails to meet the requirement, the evaluand [or evaluatee] fails overall" (Davidson, 2005b, p. 241). Soft hurdles, however, are an overall requirement for entry into a high-rating category, and failure to clear the hurdle does not classify the evaluand or evaluatee as a failure. Instead, "it places a \textit{limit} on the maximum rating that can be achieved if an evaluand [or evaluatee] does not clear a particular soft hurdle (e.g., to get an overall "A" for a course, none of the assignments completed in the semester can be lower than a "B-")" (Davidson, 2005b, p. 247).

\textsuperscript{118} A recent example, where such a violation was 'suspected,' is the case of the work of a physicist "who claimed success in conducting low-cost fusion experiments that, if validated, could point the way to a new, cheap source of power" (Monastersky, 2006, March 9). However, even though the work was published in \textit{Science}, no independent scientist has managed to replicate the work.

\textsuperscript{119} \textit{Holistic barring} normally involves a simple visual inspection of performance across all dimensions or components, where performance across all must meet a minimum in order to 'pass'; an intrinsic process that can sometimes be made explicit.

\textsuperscript{120} \textit{Global barring} involves setting or requiring minimum levels of combined or aggregated performance across all dimensions or components (e.g., $\bar{X} \geq \sigma$ in order to 'pass' or clear the global bar).
or ‘cleared’ the high bar. Clearing the high bar essentially exempts the evaluatee or evaluand from the need to have minimum levels of performance on other desirable dimensions. However, one can also ‘grab’ the low bar, for instance on the ethicality dimension. “Now, the low bar on that dimension cannot be overridden by clearing the high bar on, for example, intellectual brilliance...so if there is evidence of plagiarism or falsification of evidence in experiments, we do not allow brilliance—even if established in cases where there is zero possibility of that performance being based on plagiarism or falsification...Furthermore, if the job description specifies other duties [i.e., other core obligations of the job], not just dimensions of desirable performance...both fund-raising and coping with the paperwork could be said to be duties for which the low bar must be cleared, even if you’re Einsteinian in your intellectual performance” (M. Scriven, personal communication, May 16, 2006).

Stepping. The stepping operation is a slightly more sophisticated variation of the weighting operation, where certain levels of performance on a criterion are given greater importance, or weight, often in increments. In the evaluation of a researcher’s performance, influence may be a criterion where performance between merit categories F-C are given a weight of 1, performance at the B category is given a weight of 1 ½, and where performance at the A category is given a weight of 2, for example. The stepping operation requires the same justification as the weighting and barring operations.

Scaling. Scaling as an evaluative operation, as opposed to a strictly measurement operation (e.g., multidimensional scaling), refers to “increasing the scale of a project or program or approach” (Scriven, 1991, p. 322) and the likely results of those increases. Like apportioning, scaling may be one of the most important and useful, albeit trickiest, operations in the evaluation of research. The scaling operation is characterized by its predictive function; for instance, will increased inputs result in increased, or better, outputs? Often, it is useful to employ a multidimensional approach for the scaling operation. In the case of

121 “Why bother with the ‘high bar’ concept, when you can get pretty close with A+?...because it’s a crucial part of the argument against treating ‘cumulative score’ as the basis for valid judgments of overall merit, in all evaluation, not just personnel, that is, against the commonly proposed quantitative analysis of qualitative judgments of merit...and this point is not restricted to such exotic cases [the Einstein case]; it is exemplified in a common and perfectly valid marking process used for exam papers or selecting candidates for research fellowships” (M. Scriven, personal communication, May 16, 2006).
122 Where the metric of merit is: A = excellent; B = good; C = satisfactory; D = poor; and F = unacceptable (see Grading).
123 Bars and steps may also be ‘fuzzy’ (e.g., confidence intervals) as well as precise (Scriven, 2005c).
124 Prediction does not require any understanding of why something is going to occur; it only requires a time-referring generalization of the type: when x, then y.
research, the relevant dimensions might be quantity, quality, and significance, for example.

Profiling. The profiling operation, in evaluation, refers to graphically exhibiting grades, not scores, on the relevant dimensions of merit. In part, its usefulness is that it avoids the “difficult” and “disputable” task of weighting, and hence the synthesis operation (Scriven, 2005d, p. 238), while in some cases it still makes ranking of two or more evaluands or values possible. Profiling is equally valuable for absolute performance profiling, where it can be used for formative, improvement purposes by identifying areas of poor or underperformance; that is, profiling can sometimes serve as a useful diagnostic tool.

Core Dimensions of Evaluation

There are five core dimensions of evaluation, which Scriven has termed ‘subevaluations’ (2005c). These five dimensions are process, outcomes, costs, comparisons, and generalizability. Each of the core dimensions requires combining both factual and value premises to arrive at essentially evaluative claims on each of the core dimensions. In some cases these can be synthesized into an overall conclusion of merit, worth, and/or significance (Davidson, 2005b; Gugiu & Persaud, 2006, April; Scriven, 2005c); normally using either the QWS or NWS operation.

While the core dimensions of evaluation are primarily aimed at application for the evaluation of programs, and to a lesser degree, policies, they are nonetheless useful in

---

125 These five core dimensions (i.e., subevaluations) as presented here have been modified from Scriven’s (2005c) KEC checkpoints 6-10.

126 Comparisons and generalizability, in particular, demonstrate that evaluation is not merely an empirical endeavor, but also a creative one. Moreover, both generalizability and costs are examples of values from the lists given in Values and Valuing and are considered core dimensions, or subevaluations, because of “(i) their virtual universal importance, (ii) the frequency with which they are omitted when they should have been included, and (iii) because they involve techniques of a relatively special kind” (Scriven, 2005c, p. 9).

127 The relevant values or criteria derived from the list given in Values and Valuing may apply across more than one of the five core dimensions. However, all of the relevant values are brought to bear on the process and outcomes evaluation dimensions. For example, research ‘productivity’ may on the one hand apply to outcomes evaluation, but, on the other hand also apply to costs (e.g., productivity in terms of what it cost) and comparisons (e.g., productivity in terms of what could have been produced); although, this normally requires modification (i.e., no criterion should be ‘counted’ more than once) of the criterion.
some cases for the evaluation of products, performances, personnel, proposals, and portfolios. For the evaluation of researchers and research it may or may not be necessary to conduct all five of the core subevaluations. In evaluating a research artifact or body of work—even as part of performance or personnel evaluation—process evaluation may not be relevant in all cases, unless there is just cause to believe that an ethical or legal value was violated, for example, or if the research process is crucial to determining the merits or value of the research (e.g., Were the instruments used valid? Was the literature review adequate?). In cases where the costs are low enough that they are unimportant, or are not at issue, the costs evaluation may not be required.128

Process Evaluation

Process evaluation is normally the assessment of everything that occurs prior to the emergence of true outcomes. In most cases, process evaluation includes the evaluation of the merit, worth, and/or significance of outputs, vision, design, planning, operation, justification (e.g., of goals), fidelity, management, activities, procedures, and so forth.129

Outputs are the tangible products that result from an evaluand’s activities (Mark, 2005), which often occur “en route” (Scriven, 2005c, p. 7) to true outcomes. In the domain of research, these types of outputs might include publications produced by a

128 Often, evaluations are subject to severe cost and time restrictions; therefore, it is often useful to employ the maxim of ‘fire at the horses first’ (Scriven, 1991). That is, when comparing multiple evaluands or evaluees (e.g., candidates for a research position), barred dimensions should be checked first, rather than evaluating them on other, often less important, dimensions. Failure to clear a bar in these types of cases is the ‘death card’ (i.e., the candidate is out of the running), which is essentially a resource saving ‘cutting’ procedure. With large numbers of evaluands or evaluees, cutting can often be done sequentially (i.e., first cut, second cut, and so on) until a reasonably sound conclusion can be reached (e.g., ‘best’ candidate).

129 Davidson (2005b) suggests that most of these process elements fall into one of three categories: content (e.g., what the evaluand consists of; that is, its basic components or design); implementation (e.g., how well or efficiently the evaluand was implemented or delivered to those who needed it); or other features (e.g., any other elements or features that make the evaluand good or bad, but that are not covered by content or implementation and are not outcomes or costs-related).
researcher or research group, patents produced through a R&D process, or a R&D product or technology, for example. These are usually included in process evaluation, however when they are truly significant outputs they should be covered in outcomes evaluation.

The fidelity aspect of process evaluation in the context of research, for instance, could consist of the execution of research as designed or agreed upon, whereas management, activities, and other procedures might be the evaluation of these elements not only within the constraints of law and ethics, but also in terms of their effectiveness and efficiency.

Process evaluation is crucial in the evaluation of R&D, particularly for formative purposes, given that a failure or breakdown in the R&D process can often result in the failure of the entire endeavor. Therefore, process evaluation is sometimes vital for avoiding disastrous results; for instance, by uncovering early warnings in an effort to avoid, avert, or reduce the worst possible outcomes. In most personnel or performance evaluations of researchers, ethical and legal values are likely to be covered in process evaluation.

Evaluation of the research process may also include a wider variety of considerations such as the adequacy of the literature review, the validity of instruments used, and so on. Moreover, process evaluation, ideally, refers to evaluation of the direct process variables that can be coupled to outcomes.
Outcomes Evaluation

Outcomes evaluation is the assessment of an evaluand’s good and bad effects. These include direct and indirect effects, intended and unintended effects, and proximal (i.e., immediate), medial (i.e., short-term/intermediate), and distal (i.e., long-term) effects. Outcomes or effects may also be singular, multiple, or hierarchical (Mathison, 2005).

Proximal outcomes or effects are frequently referred to as outputs, and usually covered under process evaluation. However, they are sometimes covered in outcomes evaluation, especially “if their role is that of an intermediate cause or intended cause of main outcomes” (Scriven, 2005c, p. 8); for instance, in the evaluation of research, a truly important discovery. By contrast, medial and distal outcomes or effects are usually time dependent, often slow occurring, and sometimes extremely difficult to observe (e.g., very small or negligible effects). For instance, citations to published research are often considered as research outcomes; that is, they are regarded as a measure of scientific

---

130 Combined, the process and outcomes evaluations are usually the most important of the core dimensions for determining merit; although this often requires at least over-bar performance on the costs, comparisons, and generalizability dimensions as well.
131 A key task is side-effect searching, and “finding outcomes cannot be done by hypothesis-testing methodology, because often the most important effects are unanticipated ones” (Scriven, 2005c, p. 8).
132 Outcomes evaluation is conceptually equivalent to internal validity (Mohr, 1995), whereas generalizability evaluation more closely approximates external validity (Bracht & Glass, 1968; Campbell & Stanley, 1963).
133 Effects and outcomes are sometimes used synonymously (Davidson, 2005b). An effect is that which is produced by a cause, and which had its beginning from some other thing (Locke, 1690/1975); or, more simply, an effect or outcome is change or lack of change caused by an evaluand. It is also worth distinguishing between side-effects, which effect the target population, and side-impacts, on non-targeted populations, given that many of the most important outcomes are often unintended or unanticipated (Scriven, 2005c). Many times these attributes are not immediately evident and emerge during the course of the evaluation.
134 Normally, outcomes are operationalized as enduring changes. For instance, at an individual level these changes may be in knowledge, skills, and/or abilities, while at an organizational level they may be changes in policies, practices, or capacity, or at a community level they might include changes in employment rates, school achievement, or recycling, and at the policy or government level they might include changes in laws and regulations (Mathison, 2005).
135 Effect and outcomes normally, but not always, occur in a temporal order.
impact or quality (Aksnes, 2005).\footnote{Citation counts might not be correctly classified as a \textit{criterion}; however, they are frequently used as an auxiliary \textit{indicator} of research outcomes and serve as a useful illustration of outcomes evaluation.} However, citations do not normally begin to appear until the first year following publication and tend to peak around the second or third year in the ‘average’ case.\footnote{Often, citations are spurious measures of research quality. They discriminate against those working on the leading edge, those who work in areas with few others, young researchers, and those working on unfashionable topics, for example. Furthermore, as Scriven notes, citations also “invalidly discriminate in favor of those who invent new terms (‘summative evaluation’ has an astronomical citation index, but it was just a handy term, not the special theory of relativity)” (1991, p. 81). He also stated that “the most plausible use [of citations] is in evaluating the \textit{significance} of a particular journal article \textit{within} a field...its significance in this sense is very loosely related to merit” (Scriven, 1991, p. 82).} For truly important research or researchers, references (i.e., citations) to their work often continue, in some cases, for decades following publication of the original work. Normally, but not always, medial outcomes or effects are usually a precondition for distal outcomes or effects, particularly to the degree to which research has made a more permanent contribution to knowledge (Moed, Burger, Frankfort, & van Raan, 1985).

However, not all outcomes are easily demonstrated; particularly the majority of research outcomes. For example, research in physics is often aimed at a better understanding of the laws of nature that govern the behavior of matter and energy. Research into materials that are superconducting at low temperatures might result in eventual outcomes of knowledge about synthesis of materials that are superconducting at room temperature. From this emerge other outcomes, for example, new classes of electronic devices and high-efficiency motors and power-transmission systems. However, these outcomes might not occur for several years, if at all. Also, this research might demonstrate that such materials cannot be made and the result would be a savings from the pursuit of such outcomes in the future (COSEUP, 2001).

The totality of outcomes and effects is often simply referred to as impact. In the case of research this would include the impact on knowledge, society, policy, or practice, among others. The major difficulty of outcomes evaluation is not determining whether
there are effects, or the breadth or depth of them, but rather determining their merit, worth, and/or significance.\textsuperscript{138, 139} Moreover, these types of conclusions are often best supported by probative inference (Scriven, 2007a). That is, an inference that is supported—weakly or strongly—by the balance of evidence, as opposed to the usual probabilistic or frequentist notions of inference.

However, in the case of evaluating research, particularly large-scale research initiatives, programs, or policies, measurable indices of impact continue to elude evaluators (Kane & Trochim, 2006, November; Lal, 2006, November). Often, research impacts are estimated by applying econometric models to test the strength of theoretical relationships between investments in R&D and growth in gross domestic product (GDP), for example (Feldman & Kelley, 2001; Hall, Link, & Scott, 2002; Ruegg & Feller, 2002),\textsuperscript{140} or various employment or quality of life indicators (i.e., social benefits versus purely economic benefits).\textsuperscript{141}

Costs Evaluation

Even if an evaluand generates extremely valuable outcomes, these must be considered in terms of what it cost to produce them (i.e., resource economy values) and informed decisions require information on both (Persaud, 2005).\textsuperscript{142} Usually, costs are

\footnotesize{\textsuperscript{138} Although Cohen (1988) provided rules of thumb for characterizing what effect sizes are small (e.g., }d\textsuperscript{ }\approx0.20\text{ or }r\textsuperscript{ }\approx0.10,\text{} medium (e.g., }d\textsuperscript{ }\approx0.50\text{ or }r\textsuperscript{ }\approx0.30,\text{} or large (e.g., }d\textsuperscript{ }\approx0.80\text{ or }r\textsuperscript{ }\approx0.50,\text{ as regarded his impressions of the typicality of effects in the social sciences, he also emphasized that the interpretation of effects requires thinking in terms of a specific area of inquiry. Furthermore, the evaluation of effect sizes inherently requires an explicit assessment regarding the practical importance of the effects.

\textsuperscript{139} For independent samples or groups, Cohen's }d\text{ is defined as: }\frac{\bar{X}_1 - \bar{X}_2}{\sigma_{pooled}}\text{ or }\frac{\bar{X}_1 - \bar{X}_2}{\sigma}.


\textsuperscript{141} Equally problematic is the fact that many effects are }idiopathic\text{ (i.e., arising spontaneously or from an obscure or unknown cause).

\textsuperscript{142} Persaud (2005) also notes that costs evaluation is infrequent, and where it is conducted, it is "often poorly done because many evaluators lack the necessary technical skills" (p. 82).}
considered as the negative utility (i.e., disutility) “incurred in the making or getting of something” (Scriven, 1991, p. 104). Costs evaluation includes assessment of monetary and non-monetary costs, direct and indirect costs, and actual and opportunity costs. This dimension considers not only classical costs-benefits, cost-effectiveness, cost-utility, cost-feasibility, return on investment analyses, and financial ratio analyses (Yates, 1996), if applicable, but also the assessment of costs which are rarely coverable by money such as stress, political and personal capital, environmental impact, and costs associated with externalities. These costs can normally be classified on three dimensions: types of costs; costs to whom; and costs when (Davidson, 2005b; Scriven, 1991).

Opportunity costs—the value given up by selecting one of several, sometimes infinite, mutually exclusive alternatives—is of central importance in the evaluation of research. In efforts to optimize scarce resources, research funders must make decisions as to which proposals should receive funding, and at what amount, as well as those proposals which are to remain unfunded (Scriven, 2006f). These opportunity costs also arise in setting research priorities or agendas, given that not everything can receive priority. These decisions frequently involve risk-related values also, such as the costs of failure versus the costs of success, and as Gilovich (1991) noted, there is a widely held belief in the ‘hot hand’ phenomenon; that is, the perception that “success breeds success” and “failure breeds failure” (p. 11).

A range of analytic techniques has been used to evaluate research investments. Economic evaluation of such investments has mainly involved cost-benefit analysis, although production function models and simulation studies have also been used, as well as internal rates of return (Link, 1996). As with the evaluation of other investments, cost-benefit analysis has been applied to evaluate the economic merits of public investment in different research areas but the technique has not been used regularly in research management. In 2001, the National Research Council (NRC) completed a congressionally mandated assessment of the benefits and costs of the Department of Energy’s (DOE) fossil energy and energy efficiency R&D programs; Energy Research at DOE: Was It Worth It? Congress followed this retrospective study by directing DOE to request the NRC to develop a methodology for assessing prospective benefits of research agendas.

The costs evaluation dimension is one of the most important of the core dimensions for determining worth.

In some respects, these beliefs are merely misrepresentations of random events.
What does this mean for costs evaluation? It means that funders of research and proposal review committees have a tendency to believe that prior performance—success or failure—is a rational basis for making funding or priority decisions, which ultimately result in opportunity costs (i.e., displacement of what could have been funded or prioritized). In the context of risk, this simply means that the risk-aversive will have a tendency to fund previously funded researchers or research groups that were successful. While the risk-aversive might pursue a 10:1 ratio of success to failure, it is also possible that the benefits produced by the ten successes are very small or trivial in the pursuit of risk avoidance. However, for the risk taker, a 1:10 ratio of success to failure can produce substantial benefits which far outweigh the costs of ten failures if one success is of significant value or importance.

Comparative Evaluation

Comparative evaluation, or comparisons, contrasts the evaluand or evaluatee with alternatives or ‘critical competitors’ (Scriven, 2005c). That is, the evaluand is compared with alternative ways for getting the same or similar benefits from about the same resources (Coryn, 2006b). Usually this requires, at a minimum, comparison with an economical alternative (i.e., ‘el cheapo’) that is equally effective, as well as a costlier alternative (i.e., ‘el magnifico’), that although more expensive, produces much greater benefits. These comparisons should be made within the constraints of resources available to the evaluand. Comparative evaluation is often done, implicitly or explicitly, as

---

146 In some cases, these values might apply to the generalizability dimension.

147 Critical competitors are similar in most respects to ‘alternative scenarios’ (Coryn, 2006b; Scriven, 2005c), and usually apply to both comparative and generalizability evaluation. In some cases these comparisons can provide an opportunity to avoid undesirable outcomes, comparatively high costs-benefits ratios, and general quality improvements, among others (Coryn, 2006b).
part of the process of setting national research priorities or agendas as well as in the review of research proposals for funding.\textsuperscript{148}

**Generalizability Evaluation**

Generalizability evaluation is more or less equivalent to the concept of external validity (Bracht & Glass, 1968; Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002); that is, the extent to which the evaluand, or some component or aspect of the evaluand, can be generalized to another set of conditions with similar results.\textsuperscript{149} This includes generalizations to other climates (e.g., social, political, physical), other staff or personnel, on a larger or smaller scale (i.e., scaling), other recipients (i.e., population validity), as well as to the exportability, transferability, transportability, sustainability, longevity, durability, and resilience of the evaluand.\textsuperscript{150} Combined, these are what Scriven refers to as “dimensions of generalization” (2005c, p. 9).\textsuperscript{151}

The basis for these inferences is a thorough knowledge of the evaluand, the broader context, and that they (generalizations) can be claimed ‘beyond a reasonable doubt’ (BRD), not necessarily through experimental or even empirical demonstration (Sciven, 2006e). In essence, generalizability evaluation requires making predictions about outcomes in alternative scenarios, and “although risky, this sometimes generates the greatest contribution of evaluation to improvement of the world” (Sciven, 2005c, p.

\textsuperscript{148} See Chapter III for a more detailed discussion of comparative evaluation.

\textsuperscript{149} Cronbach (1966) contended that generalizations of this type will have to be stated with several qualifications, such as: “With subject matter of this nature, inductive experience of this type, in this amount, produces this pattern of responses, in pupils at this level of development” (p. 77).

\textsuperscript{150} These are similar in certain respects to the conceptions of population validity and ecological validity (Bracht & Glass, 1968), neither of which are entirely independent.

\textsuperscript{151} The generalizability dimension is normally the most important of the core dimensions for determining significance.
Generalizability evaluation, as described above, requires slightly different inferences for science and technology (S&T) and R&D than it does for a researcher or an instance or piece of research in most cases (i.e., the generalizability of a S&T or R&D program to another set of conditions, the versatility or utility of a researcher after transportation to another institution, department, or center, or the generalizability of research results, for instance)—though the overarching rationale is the same (i.e., the extent to which the evaluand or aspects of the evaluand can be generalized to another set of conditions with similar results).

Fields of Evaluation

There are now more than a dozen recognized branches or fields of evaluation, of which the “Big Seven” (Scriven, 2005d, p. 237) are said to be the evaluation of products, performances, personnel, programs, policies, proposals, and portfolios. The Big Seven have a “long history of practice” but a “shorter history of methodological discussion” (Scriven, 1991, pp. 165-166). In most cases, the evaluation of research or researchers is an example of performance and product evaluation; the performance of a nation, institution, research group, or researcher and the (research) products of them, although it often involves elements of personnel and proposal evaluation. In any case, the general working logic of evaluation (Fournier, 2005) is analogous for the Big Seven.

---

152 In research, generalizations are typically aimed at populations through extrapolation. However, these types of generalization often involve tenuous inductive and imaginative leaps.

153 However, as Cronbach (1975) observed, generalizations decay and what at one time describes the social situation well might later be valid only as history.

154 Intradisciplinary and metaevaluation (Scriven, 1969)—the evaluation of evaluations—are also often given status as evaluation-specific fields. The United States Supreme Court is an example of metaevaluation where decisions by Appellate Courts are evaluated.

155 Product evaluation in the domain of research is interpreted very broadly; for example, a new pedagogical process might be the product of a R&D process.

156 Fournier has claimed that evaluation has a basic logic as shown in these four steps (1995, p 16): (1) establishing criteria of merit—on what dimensions must the evaluand do well? (2) constructing standards—how well should the evaluand perform? (3) measuring performance and comparing with standards—how well did the
Each of the Seven is strongly tied to evaluating research and can be viewed mostly in terms of their increasing levels of molarity. For instance, evaluations of research personnel and research proposals are instances of evaluation at a molecular level, while evaluation of research programs, institutions, or fields is more consistent with evaluation at a macro or molar level, whereas research performance or portfolios might be viewed as micro-level evaluation (depending upon whose performance or portfolio; for instance, a single researcher versus a research department or institution). Moreover, molecular-level evaluations of research personnel or products, for example, are often used as a basis for macro- or molar-level policy or program evaluations.

Program Evaluation

Of the Big Seven, program evaluation\(^{157}\) receives the most attention and has the most well-developed principles, procedures, and practices.\(^{158}\) While most historians of evaluation credit the emergence of systematic program evaluation to the Johnson and Kennedy administrations in the 1960s, Madaus and Stufflebeam (2000), however, claim that it originated in 19\(^{th}\) century Great Britain, where attempts to reform education, law, hospitals, orphanages, and public health were evaluated by government-appointed commissions.

Whether program evaluation emerged in Great Britain in the mid-1800s or in the United States in the mid-1900s, contemporary program evaluation has flourished with the development of evaluation-specific methods, models, theories, approaches, and so

---

\(^{157}\) Programs are loosely defined as “systems under which action may be taken toward a goal” (Schröter, 2006, November).

\(^{158}\) The importance of product and personnel evaluation within program evaluation has yet to be taken seriously. However, product evaluation is probably the “oldest practice within evaluation, although explicit discussions of its methodology has not received much attention” (Scriven, 1994c, p. 46).
forth. However, evaluation has yet to develop what could be considered an explicit metatheory, although Scriven (1982, 1991, 1983, 2005, October) has laid the foundations with the transdisciplinary model (Coryn & Hattie, 2006; Scriven, 1991). Evaluations of research programs, R&D centers, institutions, and so on, have a relatively long history. However, most have serious weaknesses in their understanding and application of the logic of evaluation.

Personnel Evaluation

As noted earlier in this chapter, in the field of personnel evaluation, the earliest documented efforts were those of the Chinese and Egyptian dynasties more than 2,000 years ago (Frechtling, 2002; Lu & Xie, 2005; Scriven, 1991). These personnel evaluations were used to draw the best talent into government and perfected by the Chinese during the Tang dynasty (618-907 AD). In the early part of the 20th century, modern personnel evaluation began to be formalized with Taylor's (1911/2003) scientific management, which emphasized efficiency.

More recently, these practices have grown to include research-based teacher evaluation (RBTE) (Peterson, Kromrey, & Smith, 1994; Schwab, 1990), personality taxonomies such as the 'Big Five' (Barrick & Mount, 1991) and other batteries of psychological tests (Cronbach & Gleser, 1957), research-based predictors of future job performance or 'organizational fit' (Schmidt & Hunter, 1998), various interview techniques (Eder & Harris, 1999), and large-scale assessment centers (Klimoski &

---

159 Personnel evaluation typically involves an assessment of job-related skills, and is usually more complex than, for example, product evaluation and in some cases may include multiple performance evaluations.
160 For example, asking inappropriate or irrelevant questions. Normally, personnel interviews involve questions of "what can the candidate do?" "what will the candidate do?" and "chemistry" (e.g., fit with the organization). The most common interviewing methodology is the past behavioral interview (PBI). A PBI includes questions about a person's experiences performing certain activities—such as managing deadlines or resolving conflicts—but does not include personal questions. This form of interview has become the
Brickner, 1987). Personnel evaluation of researchers is often based on the faculty evaluation model, which emphasizes teaching, research, and service (Arreola, 2000; Scriven, 1991). Although the common approaches to personnel evaluation have generated some major practical innovations, it is still far from ideal and a long way from the principles of systematic and objective evaluation.

Performance Evaluation

As a subfield of evaluation, performance evaluation is slightly more developed than personnel evaluation, particularly as it relates to testing and assessment. In 1792, modern psychometrics began to develop when William Farish replaced traditional qualitative assessments of student performance with quantitative markings of correct and incorrect answers, which allowed for ranking examinees as well as averaging and aggregating of test scores (Madaus & Stufflebeam, 2000).

While performance evaluation has primarily come to refer to student performance on tests, whether for individual students, a class, a school, or a district it also has a long history of application in the judging of athletic prowess and performance (Scriven, 1991), for example, in Olympic and other sporting competitions (Weekley & Gier, 1989). Performance evaluation also serves as a basis for assessing the uniformity accepted best practice over the past thirty years, and some claim that the PBI is a good predictor of performance (Menkes, 2005). It can explain about 25% of the variation in performance among employees.

Most modern personnel evaluation practices and procedures are drawn from research conducted in human resources (HR) and industrial and organizational (IO) psychology.

Performance evaluation is the evaluation of a particular achievement, in the form of an output or process, and for example, include “a student’s performance on a test (or across a term) and a gymnast’s routine on a particular apparatus” (Scriven, 1991, p. 256).

In professional figure skating, for example, technical marks are awarded individually for each skating element (the number and type of elements in a skating program depends on the event and on the level of competition). Competitive programs are constrained to have a set number of elements. Each element is first judged by a technical specialist who identifies the specific element. The decision of the technical specialist determines the base value of the element. A panel of twelve judges then awards a mark for grade of execution (GOE) that is an integer from -3 to +3. The GOE mark is then translated into a value by using a table of values. The GOE value from the twelve judges is then averaged by randomly selecting
of performance of a researcher or R&D center, for example, in reference to constant high quality. Often, these past performances serve as a basis for allocating resources, such as research funding, and play an important role in evaluating research proposals.

Product Evaluation

Product evaluation,\(^{164}\) as discussed earlier in this chapter, has a long history of practice dating back several millennia to the product evaluations conducted by early craftsmen, artisans, guilds, and professional societies.\(^{165}\) In the last 50 years, however, product evaluation has become considerably more extensive, public, and sophisticated (Scriven, 1994c). Despite this long history of practice, contemporary evaluation of products, particularly consumer products, on which lives, and the quality of lives, often depend, lack an adequate understanding of evaluation-specific logic, among other shortcomings.

While examples of high quality product evaluations abound, such as the evaluation of drugs by the Food and Drug Administration (FDA) and the automobile crash tests conducted by the Insurance Institute for Highway Safety (IIHS),\(^{166}\) as a subfield of evaluation, product evaluation, in general, still suffers from goal-orientedness, fallacies of technicism,\(^ {167}\) aestheticism, and irrelevant and biased expertise (Scriven, nine judges, discarding the high and low value, and averaging the remaining seven. The average value is then added (or subtracted) from the base value to arrive at the value for the element.

\(^{164}\) Product evaluation is generally taken to mean “the evaluation of functional artifacts, but it can also be taken to include the evaluation of output from students, such as essays” (Scriven, 1991, p. 280); however, these would normally be classified under performance evaluation (see the Process Evaluation and Outcomes Evaluation subsections of this section).

\(^{165}\) While research products (i.e., a particular instance or piece of research) can usually be evaluated independently of the researcher (i.e., evaluation of the researcher’s research), the reciprocal (i.e., evaluation of the researcher excluding their research) is almost never true.

\(^{166}\) Recently, the IIHS has discontinued its front crash tests and is initiating a new approach involving evaluations based on manufacturers’ own frontal tests of vehicles meeting requirements established by the IIHS. Manufacturers will provide detailed information from their offset tests, including video, and the IIHS will assess this information, assign ratings, and conduct audit tests to verify manufacturers’ results (Insurance Institute for Highway Safety, 2006, March).

\(^{167}\) Technicism is the valuing of pragmatically meaningless technical specifications.
As will be seen, these problems also plague the evaluation of research products as well, particularly in the application of implicit quality criteria for determining the merits of a research product; for example, that research is good if it is conducted within a certain philosophical or epistemological perspective.

Policy Evaluation

Policy evaluation emerged almost simultaneously with program evaluation in the United States in the 1960s and 1970s from "operations research, microeconomics, organizational theory, public administration, social psychology, and the increasing interest in the role of law in public policy" (Scriven, 1991, p. 267). While policy evaluation shares a number of characteristics with program evaluation (e.g., working logic, methodology), the evaluation of policy is normally either retrospective evaluation of implemented policy or prospective evaluation of possible policy or comparisons of alternative policies.

Unlike program and other subfields of evaluation however, policy evaluation, particularly of the prospective type, frequently addresses the "now what?" types of evaluative questions, which often require sophisticated micro-simulations or formal modeling techniques for "determining the merit, value and worth of...whatever governments choose to do or not to do" (Risley, 2004, February) or the evaluation of alternative government policies or decisions in order to arrive at the best, or a good, policy (Nagel, 1990). Its role in evaluating research is two-fold. First, it is the evaluation of research policy in matters such as priorities and agendas, funding and allocation and

---

168 The apex of modern product evaluation is the evaluation of automobiles, although there are still serious errors and room for improvement (Scriven 1994b).

169 Policy evaluation, often referred to as policy analysis, is the evaluation of policies, plans, and sometimes proposals and possibilities, and the role of the policy evaluator, or policy analyst, "is somewhat different than that of the evaluator—it is scenario evaluation, that is, evaluation of alternative possible futures" (Scriven, 1991, p. 267).
distribution of those funds, for example (Coryn, 2007, January). Second, it is policy related specifically to large-scale evaluations of research such as those used for evaluating government-funded research.

**Portfolio Evaluation**

Portfolio evaluation is a newly emerging subfield of evaluation and it remains, in large part, underdeveloped. While in the business world a portfolio normally refers to a portfolio of investments, in the arts, architecture, music, teaching, technology, and most forms of research, for example, it typically refers to a body or selection of professional achievement. Around the 1980s, portfolio evaluation began to be used to assess students in areas such as writing, mathematics, English, and history, and in some cases has replaced or supplemented performance evaluation, and testing and assessment, as indicators of student learning and achievement (Berlach, 1997). In the early 1990s, portfolio evaluation was applied to the evaluation of faculty teaching, research, and service (Arreola, 2000).

More recently, Web-based portfolios have been used for evaluating courses, curricula, and institutions (Banta, 2003). However, portfolio evaluation has yet to resolve a number of serious problems, including, how to deal with conflicting values, or criteria, such as diversity versus specificity, risk versus risk-aversion, and quantity versus quality, among others. Research portfolios frequently serve as the basis for evaluating

---

170 Portfolio evaluation takes many forms, for instance, the evaluation of an artist’s portfolio (a body of artistic work), the evaluation of an investment portfolio (e.g., the securities held by an investor), the evaluation of a teaching or research portfolio, and so forth.

171 Often the evaluation of investment portfolios employs uncertainty and probabilistic analyses in an effort to reduce risks and increase gains; frequently referred to as decision analysis (Clemen, 1996).

172 In the case of a researcher’s portfolio, for instance, does diversity or specificity constitute a ‘good’ body of work (i.e., a research portfolio)?
researcher performance, for personnel evaluations of researchers, or for national investments in science and R&D (Jordan, Hage, & Mote, 2006, November).

Proposal Evaluation

The evaluation of proposals, particularly of research proposals submitted for funding, has a relatively long tradition of practice, for example, by NSF and the National Institutes of Health (NIH) in the United States, and is characterized in part by its predictive and sorting function. These types of proposal evaluations are normally assessments of the proposer's ability to perform the prospective research successfully and whether the anticipated research outcomes are worthwhile.

Normally, these assessments are conducted by peer review panels to judge likely, or anticipated, future performance on the basis of prior performance. These types of proposal evaluations are normally applied to determine the intellectual merits and significance of the proposed research. However, as it is currently practiced, proposal evaluation has a number of serious flaws and is badly in need of study and reform. For instance, proposal review committees and referees often lack standardized or calibrated rating procedures, are frequently manipulated by special interests, are often politically controlled, are subject to conflicts of interest, are not truly blinded, sometimes just plain laissez-faire in their reviewing practices, and so forth.

173 Proposal evaluation is normally thought of as having two variations: (i) the evaluation of systematic suggestions, often in the form of plans and (ii) the evaluation of proposals submitted for funding, typically to foundations or government agencies. However, proposal abstracts are also evaluated for their merits for presentation at meetings of professional organizations (Schröter, Coryn, & Montrose, 2006).

174 NSF applies two explicit criteria in evaluating proposals: (i) intellectual merit and (ii) broader impacts (National Science Foundation, 2004).

175 NIH applies five explicit criteria in evaluating proposals: (i) significance; (ii) approach; (iii) innovation; (iv) investigators; and (v) environment (National Institutes for Health, 2004).

176 During the last decade, the process of evaluating student's dissertation and thesis proposals have also come under scrutiny (Pathirage, Haigh, Amaratunga, Baldry, & Green, 2005).

177 A third variant of proposal evaluation are the reviews conducted by institutional or other types of ethics committees or review boards to interpret and apply federal regulations, state law, and research sponsor requirements for the use of human subjects in research.
The Transdisciplinary Model

The transdisciplinary view, or model, of evaluation requires an understanding of how and why evaluation developed from a practice to a highly skilled, professional practice to a field-specific discipline, and finally to an autonomous discipline and transdiscipline, much like ethics, statistics, and measurement (Scriven, 2003). This understanding, in part, becomes known from the transdisciplinary model's three primary characteristics that make it a transdiscipline, which are: epistemological; political; and disciplinary (Scriven, 1993).

The epistemological characteristic of the transdisciplinary model is one drawn from an objectivist view of evaluation. This is a paradoxical notion, and despite the various meanings or definitions assigned to the concept by various disciplines, schools of thought, or individuals, there is ultimately a body of knowledge representative of a single reality. Objectivity is considered as the compatibility of objective propositions distinct and independent of subjective propositional attitudes or acts. The nature of a proposition is that it must be true or false, and its many forms include the axioms and formulas of the sciences and mathematics, as well as the rules and processes of logic. Therefore, the objectivist view of evaluation asserts that evaluative claims of merit, worth, and significance are possible in principle and practice, based on logic and reason, and if properly understood, objectivity.

---

178 A transdiscipline is one which is based on a distinction from primary disciplines, for instance, the conventional academic disciplines, and a class of disciplines which provides some set of tools, methods, and/or approaches for use by the primary disciplines. Transdisciplines include, among others, statistics, measurement, logic, and evaluation. "Logic, with its applied fields of the logic of the social sciences and so on, is an extremely general transdiscipline, but evaluation is probably the most general (unlike logic, it precedes language); both are much more general than measurement or statistics" (Scriven, 1993, p. 9).

179 See the Toward a Discipline section of this chapter.

180 A version of this section appeared in the Journal of MultiDisciplinary Evaluation (Coryn & Hattie, 2006).

181 Moreover, objectivist evaluation is premised on "the theory that moral good is objective and independent of personal or human feelings" (Stufflebeam, 2005, p. 62).
The political characteristic of the transdisciplinary model is that it is characterized by a consumer-oriented view. That is, the rationale or justification of the existence of a program, policy, or product is to serve the needs of consumers. In the transdisciplinary view, evaluation affords those consumers "the same primacy in evaluation" and therefore the main function of evaluation is "the determination of the merit or worth [or significance]...[of a program, policy, or product]...in terms of how effectively and efficiently they are serving those affected, particularly those receiving, or who should be receiving, the services provided and those who pay for [them]" (Scriven, 1993, p. 9). Although the consumer-oriented view is grounded in a deeply reasoned view of ethics and the common good, Stufflebeam (2001a), in his taxonomy and analysis of evaluation models and approaches, described consumer-oriented evaluation as "extremely difficult" and that it requires "a highly competent and credible expert" (pp. 59-60).

The disciplinary characteristic of the transdisciplinary model, similar to statistics, ethics, and logic, is that evaluation is a discipline that can be characterized by its study and improvement of certain tools for application between and within other disciplines (Scriven, 1991, 2005a). The disciplinary characteristic of the transdisciplinary view of evaluation can be separated into three component parts: disciplines (e.g., arts, humanities, social sciences, technology, natural sciences); fields of evaluation (i.e., product, performance, personnel, program, policy, proposal, and portfolio); and fields of application (e.g., education, health, human services).

---

182 This particular view of evaluation is often referred to as consumer-oriented, consumer-based, or needs-based evaluation; although these differ slightly in their meaning.

183 Moreover, as the importance of evaluation as a consumer service, typically summative, was stressed, "some theorists...began to talk as if this was the essential duty of evaluation. Consumers, by and large, have no interest in whether the program or product designer's goals have been met, and only a secondary interest in improving the program (i.e., formative evaluation), being mainly interested in whether their own needs are met" (Scriven, 2001, p. 27).
Graphically, the disciplinary elements of the transdisciplinary model can be represented by spatial planes in three dimensions, with each plane representing one of the three disciplinary components. As illustrated in Figure 2, the rear plane on the $x$ and $y$ axes represents disciplines, the vertical plane on the $y$ and $z$ axes represents fields of evaluation, and the horizontal plane on the $x$ and $z$ axes represents fields of application. Any particular evaluation can then be located as a point or volume (e.g., a cube) in this three-dimensional space.

![Diagram of the Transdisciplinary Model](image)

Figure 2. The Transdisciplinary Model

Conceptually, the transdisciplinary model is a useful tool in the evaluation of researchers and their research, as well as for other types of evaluand's or evaluatees. For instance, in evaluating a molecular biologist's research, each of the elements of the transdisciplinary model can be spatially mapped, such as illustrated in Figure 3, where the molecular biologist's research would be classified as a natural science discipline $(x, y)$, evaluated as a product $(y, z)$, and where the biologist's research is applied in the field of health $(x, z)$, for example. This is by no means a sophisticated or exact procedure, such
as the plotting of points using a three-dimensional Cartesian coordinate system, but rather serves as a practical, illustrative representation of the nature of the evaluand or evaluatee and the evaluation thereof.

Figure 3. Spatial Map of a Discipline, Field of Evaluation, and Field of Application within the Transdisciplinary Model

Additionally, using Scriven’s analogy of the “House of Evaluation” from The Country of the Mind in the Evaluation Thesaurus (1991, p. 13), the reasoning and logic of the transdisciplinary model can be extended somewhat further, and the placement of the disciplinary elements of the model (i.e., disciplines, fields of evaluation, and fields of application) clarified and their interrelatedness revealed by expansion of the allegory to construct the framework for the house, situated somewhere near logic and ethics in the geographical landscape of the “country of the mind.” As illustrated in Figure 4, the metaphorical floors of the house include, but are not limited to: (i) the ground floor (i.e., fields of application plane on the x and z axes), which represents applied work, above which are floors representing, or dedicated to (ii) the development of instruments; (iii)
methods; and finally (iv) theory on the top floor; with (v) metatheory hidden in the attic. Moving upward, toward the floors occupied by theory and metatheory, each gets increasingly smaller.

Figure 4. The Transdisciplinary Model and the House of Evaluation

Reasons and Motives for Evaluating Research

The evaluation of research serves numerous purposes, although there have been extensive debates, and in general, an overwhelming lack of consensus as to what these reasons and motives are or should be (e.g., Aksnes, 2005; Moed, 2005; Rousseau, 2004, 184 Evaluation-specific theories are often local theories, about a particular field or subfield of evaluation, for instance; that is, theories about program, performance, or personnel evaluation. However, general theories are, for the most part, lacking. Evaluation models and approaches, for instance, empowerment (Fetterman, 2001), utilization-focused (Patton, 1997), the CIPP model (Stufflebeam, 2004), and so on, do not qualify as theories in the true sense; these are normally "metaphors for, conceptualizations of, or procedural paradigms for evaluation" and "the latter come closest to being theories in the usual sense, the others are nearer to metatheories" (Scriven, 1991, p. 156).
October; Russell & Rousseau, 2002; van Raan, 2005).\textsuperscript{185,186} In part, this disagreement can be attributed to the larger context in which the evaluation of research takes place.\textsuperscript{187} In most cases the evaluation of a nation's research serves vastly different purposes than evaluation conducted by a department or research group evaluating candidates for a research position, tenure, promotion, or demotion, or than the evaluations conducted by a journal editor or peer reviewer assessing a paper's merits for publication.\textsuperscript{188}

There are essentially five fundamental purposes for evaluating research, although there is some overlap, which can be broadly classified as: accountability and efficiency; resource allocation; improvement; synthesis; and decision making.\textsuperscript{189,190} With the exception of improvement, most research evaluations are summative, and in some cases synthesis is done for ascriptive rather than summative purposes. Excluding synthesis, and as mentioned previously in this chapter, if there is a single word to describe these purposes it is 'governance' (Frederiksen, Hannson, & Wennberg, 2003). Governance is a somewhat ambiguous term for social regulatory processes that directly or indirectly implicate the political system; it is analogous to the psychologists' and sociologists' term 'social control' (Hannson, 2006).

\textsuperscript{185} As recently stated by a member of the AEA's RTD TIG: “research that isn’t funded isn’t worth evaluating” (anonymous, personal communication, October 28, 2005), with the notion being that only federally-funded R&D is ‘worth’ evaluating. While this seems to be a narrow view, it is in fact one shared by a large community of evaluators, many whom are members of WREN and responsible for evaluating DOD, DOE, National Aeronautics and Space Administration (NASA), NIH, and NSF research in the United States. However, Darwin and Einstein's research was never funded but have probably had a greater impact than any that has come since.

\textsuperscript{186} Other proposed purposes include, for example, learning and auditing of evidence-based policy and practice (Cousins, 2006, April).

\textsuperscript{187} The contextual dilemma is essentially a unit of analysis problem (e.g., macro-level evaluation of a nation's research, meso-level evaluation of a research institution, group, or department, or micro-level evaluation of a single researcher and/or their research).

\textsuperscript{188} No matter the purpose, the evaluation of research always involves one or more of the three basic evaluative predicates: merit, worth, and/or significance (see the Logic of Evaluation section of this chapter).

\textsuperscript{189} These purposes serve the complete tripartite taxonomy of evaluative purposes; formative, summative, and ascriptive (see The Tripartite Taxonomy subsection of this chapter).

\textsuperscript{190} More generally, the evaluation of research is on the one hand, the evaluation of future performance (i.e., prediction) and on the other, the evaluation of achievement to date.
In any case and whatever the purpose, the evaluation of research has been called a priori or a posteriori (Weinberg, 1963, 1989). In the first instance research is evaluated prospectively, often referred to as ex ante evaluation (Meyer-Krahmer & Reiss, 1992), for predicting future performance, normally on the basis of prior performance. In the second instance research is evaluated retrospectively, often referred to as ex post evaluation (Campbell & Felderer, 1997), after it has been completed. Ex ante evaluation of research is normally used for awarding research funding for proposed research, whereas ex post evaluation of research is applied for determining the merits or significance of completed research, for instance, in awarding Nobel Prizes.\textsuperscript{191}

In evaluating researchers and their research, accountability and efficiency, priority setting, resource allocation, synthesis, and decision making are primarily summative endeavors, although in some cases they can be done for formative, ascriptive, or less frequently, proformative (Coryn, 2007a; Scriven, 2006a) purposes.\textsuperscript{192} Improvement, however, is an entirely formative procedure in most cases, although it often occurs as a result of summative evaluation.\textsuperscript{193}

\textsuperscript{191} Nobel Prizes are widely regarded as the most prestigious awards given for intellectual achievement in the world (Crawford, 1998). They are recognized by virtually every scientist, and they are also among the few prizes known by name to many ordinary citizens. The only international prizes that approach them in importance are those awarded in the Olympic Games. Part of the Nobel Prize’s prestige stems from the serious research that goes into the selection of the prizewinners, which is shrouded in secrecy (Crawford, 1990). Several thousand people are involved in the committees’ efforts to determine the originality and importance of each nominee’s contributions, with outside experts frequently being called in during the process. The general principles governing awards were laid down by Alfred Nobel in his will. These statutory rules have on the whole remained unchanged but have been somewhat modified in application. For example, Nobel’s stipulation that the prizes be awarded for achievements made during ‘the preceding year’ was obviously unworkable in regard to most scientists and writers, the true significance of whose discoveries, research, or writings might not be generally apparent for several years (Lemme!, 2000).

\textsuperscript{192} Proformative evaluation “is motivated, like formative, by the intention to improve something that is still developing, but unlike formative, the improvement is only possible by taking action, hence proactive [italics added] instead of reactive, [italics added] hence both, hence proformative” (M. Scriven, personal communication, March 9, 2006). Proformative evaluation first appeared in \textit{The Great Enigma: An Evaluation Design Puzzle} (Scriven, 2006a).

\textsuperscript{193} The case where summative can result in formative evaluation is illustrated in the example in the \textit{The Tripartite Taxonomy} subsection of this chapter. Moreover, in many large-scale national evaluation systems, where the primary purpose is the allocation of resources, researchers are often forced to improve or else risk their livelihood.
Accountability and Efficiency

As a purpose for evaluating research, particularly publicly-funded research, accountability and efficiency is the responsibility for the justification of expenditures, decisions, or the results of research efforts. Accountability often requires some measure of cost-effectiveness, where cost-effectiveness is taken to be more than explanations of how financial resources were spent, but also justifications in the results produced from these expenditures. There is considerable variation in who is required to answer to whom, concerning what, through what means, and with what consequences. Economic, social, and other benefits, often referred to as impacts, are normally subsumed under accountability.

While accountability is most often considered a purpose for the evaluation of a nation’s research or its expenditures of taxpayer monies on research initiatives or agendas, it is equally applicable to research institutions, groups, departments, or individuals; that is, they are equally accountable for justifying expenditures, decisions, or the results of their research efforts. This can also be extended to include accountability for who is tenured, promoted, demoted, hired, or fired by a research institution, group, or department, for example.\footnote{There are “wholly incompetent faculty members, who impose a huge cost on students and/or taxpayers, colleagues, and those who need their job, and could do it much better” (Scriven, 1991, pp. 161-162), the identification of whom is certainly justifiable as a form of accountability.} At the personnel level, accountability serves to justify cost to students, taxpayers, colleagues, and others in the selection of researchers. In practice, however, many systems of accountability are subject to several forms of corruption and “hence are likely to reduce the sense of responsibility for and quality of performance” (Rogers, 2005, p. 2).
Resource Allocation

Resource allocation, or apportionment, is often an explicit, and in some cases implicit, purpose for the evaluation of researchers and their research. Conceptually, resource allocation involves matters such as national priority setting which normally includes the distribution of research funding (Coryn, 2007, January; Scriven, 2006f).

Resource allocation may be one of the most important purposes underlying the evaluation of research, and not entirely unrelated to accountability. Ultimately, investments in research are like other types of investments, more uncertain, but conceptually similar (Scherer, 1967). However, these allocations frequently involve a great deal of trial and error. In strategic planning, a resource-allocation decision is a plan for using available resources, especially human resources in the near term, to achieve goals for the future. It is the process of allocating resources among various projects, units, or alternatives.

A typical allocation plan has two parts. First, there is the basic allocation decision and second there are contingency mechanisms. The basic allocation decision is the choice of which items to fund in the allocation plan, and what level of funding each should receive, and which to leave unfunded. That is, resources are allocated to some, not to others. There are two contingency mechanisms. There is a priority ranking of items excluded from the plan, showing which items to fund if more resources should become available and there is a priority ranking of some items included in the plan, showing which items should be sacrificed if total funding must be reduced.

All decision makers have to work within a world where resources are scarce in comparison with alternatives for their use. Those responsible for the allocation of funds to competing lines of research are no exception to this rule of constrained decision

---

195 See the Costs Evaluation and Comparative Evaluation subsections of this chapter.
making, and certain characteristics of research make it particularly difficult to decide on the best distribution of resources. The most salient of these characteristics is that the net benefit from any line of research is, by its very nature, uncertain, since there is no sure way of predicting whether a particular researcher or group of researchers will be able to produce research of a significant value.

Improvement

Since Scriven’s introduction of the term ‘formative evaluation’ in 1967, improvement has been recognized as a fundamental purpose for many evaluative endeavors. As an explicit enterprise, however, evaluation for improvement is a relatively new and often ignored purpose for the evaluation of researchers and their research.

In some countries, the intended purpose of research evaluation is to invoke an intraregional or inter-researcher competitive spirit (Saegusa, 1999a, 1999b), in order to produce general quality improvements in its researchers and their research (Swinbanks, Nathan, & Triendl, 1997), and ultimately their place in the world’s research spectrum (e.g., their international ranking). Normally, improvement is a secondary function of national-level evaluation of publicly-funded research, expected to occur as a result of competition for research monies.

However, these efforts do not always invoke an interregional competitive spirit, but rather encourages ‘game playing’ in some cases.\footnote{Game playing sometimes occurs when evaluations are strictly conducted using indicators such as amounts of external funding or number of publications. Often, researchers modify their practices (e.g., increasing numbers of publications by publishing more frequently or in lower impact journals) in order to meet these performance criteria (i.e., quotas).} While improvement of a nation’s researchers or their research is a long way from improving the quality of a manuscript submitted to a journal for publication, it is, nevertheless, an essential function of the evaluative endeavor as it is currently understood. In some parts of the world, however,
evaluating the research of one’s peers or colleagues is still viewed as an incursion upon longstanding cultural traditions, despite the potential for general quality improvements in their research:\(^{197}\)

...research assessment is an alien concept that runs directly against the grain. This is a region, after all, in which deep-rooted traditions demand respect for elders and the promotion of harmony and cooperation at the expense of individuality and competition...openly judging the quality of scientists and firing those who do not come up to mark is hard...in cultures built on Confucian and Buddhist values of respect and group harmony (Swinbanks, Nathan, & Triendl, 1997, p. 113).

Synthesis

There are some (e.g., Campbell Collaboration, 2006; Cochrane Collaboration, 2006) who view the purpose of research evaluation as a synthesis:\(^{198}\) activity, much along the lines of modern meta-analysis or systematic review (Glass, 1976; Pawson, 2006; Rosenthal, 1976),\(^{199}\) which is primarily a summative undertaking, but also a special case of ascriptive evaluation.\(^{200}\) Essentially this view sees scientific knowledge as an accumulative endeavor and uses statistical techniques to combine the results of several studies that address a set of related research hypotheses for computing an average effect size across all relevant studies is computed using a weighted mean, whereby the weights

\(^{197}\) Nevertheless, many Eastern governments and research institutions are recognizing that more creativity and innovation in their research systems may be essential to the future success of their economies, and are rapidly adopting and adapting Western techniques of research assessment in an attempt to improve the productivity and the quality of their research output (Campbell, 1997; Coryn, 2006, October, 2006, November; Frankel & Cave, 1997; Swinbanks, Nathan, & Triendl, 1997).

\(^{198}\) Synthesis as an evaluation-specific operation is presented in the Logic of Evaluation section of this chapter. The two are distinct operations, with one being the synthesis of research findings and the other being the synthesis of an evaluand’s or evaluee’s performance.

\(^{199}\) This view is predominately held in the health care industry, although it is now beginning to take hold in the education and social service domains, where systematic reviews are taken as mechanisms for informing ‘evidence-based’ practice or policy (Glasziou, Vandenbroucke, & Chalmers, 2004).

\(^{200}\) It is also a case of summative evaluation in that it involves decision making; for instance, what is the most effective treatment for certain types of leukemia?
are equal to the inverse variance of each study's effect estimator (e.g., Cohen's $d$, Hedge's $g$, Glass' $\Delta$).

Meta-analytic studies have grown in number over the last few decades and "its popularity in the social sciences and education is nothing compared to its influence in medicine, where literally hundreds of meta-analyses have been published in the past 20 years" (Glass, 1999, p. 1). Moreover, the increasing use of meta-analysis has encouraged some researchers to view their studies as making contributions to previous research and to report their results so that they can easily be incorporated (e.g., effect sizes and confidence intervals) into future meta-analysis (Cumming & Finch, 2001).

These types of evaluations of research are useful evaluative endeavors, for example, for getting to the bottom line, identifying critical competitors, and possible side-effects, among others, and are often considered the gold standard for evidence-based policy and practice (Boruch & Herman, 2006), particularly in the health disciplines. More recently, large-scale synthesis of this type can be observed by the establishment of the United States Department of Education's (USDOE) What Works

---

201 Presently, there are two large-scale providers of meta-analytic studies. The first of these is the Cochrane Collaboration, with its *Cochrane Library*, which claims to have compiled "the best available information about healthcare interventions...the evidence for and against the effectiveness and appropriateness of treatments" (Cochrane Collaboration, 2006). The second is the Campbell Collaboration, with its *Campbell Library*, which declares that it provides information for making "well-informed decisions about the effects of interventions in the social, behavioral and educational arenas", and which "prepare[s], maintain[s] and disseminate[s] systematic reviews of studies of interventions" where research merit is determined by "high quality evidence on what works" (Campbell Collaboration, 2006). More recently, a similar provider, the Institute of Education Science's (What Works Clearinghouse, was developed to focus on the effects of education interventions (Boruch & Herman, 2006).

202 Synthesis as a purpose for the evaluation of research is not attended to in this dissertation as there are already well-developed procedures in place for these types of operations.

203 For example, "is the treatment effective?"

204 For example, "are there equally effective treatments which can be given at lower costs?"

205 According to Pawson (2006) the real purpose of systematic review is to better understand program theory, so that policies can be properly targeted and developed to counter an ever-changing landscape of social problems.

206 See the *Methodological Rigor* subsection of this chapter.
Clearinghouse (WWC), which collects, screens, and identifies studies of effectiveness of educational interventions, including programs, products, practices, and policies (2007).207

Decision Making

Decision making, although summative in purpose, has been classified as a separate purpose since there are other aspects of the decision making function involved in the evaluation of researchers and their research than the usual summative issues: whether or not one has been accountable for research spending; if research is worthy of synthesis to inform policy or practice; or if research resources have been distributed justly.208 It also involves matters such as selection, prioritization, and prediction. For selective purposes, decision making involves the evaluation of proposals, whether or not for funding, research submitted for publication, research products, and research personnel.209 That is, “which research proposals receive funding, which articles get published, and which researchers...get appointed and promoted” (Frankel & Cave, 1997, p. 1).

Priority setting in research, usually at the national level, serves the purpose of answering questions such as “now what?” “how much?” and “to whom?”210 For example, “whether or not to go to the moon, and how much should go for the support of high energy physics” (Weinberg, 1989, pp. 4-5). Priority setting, while a purpose for

207 The WWC works jointly with the American Institutes for Research and the Campbell Collaboration.
208 All purposes for evaluating research involve decision making, even in the formative sense. Although formative evaluation is done to determine areas which need improvement, making improvements requires action.
209 Decision-making in matters related to research, particularly research products (e.g., findings or results), is an extremely sophisticated endeavor and not all decisions related to research are those made by funders of research (decision = fund the research or do not fund the research) or employers of researchers (decision = promote, demote, hire, or fire a researcher). For example, decisions are also made by research consumers (e.g., the public), where epidemiological research supporting the connection between secondhand smoke and cancer, for instance, may lead to an individual decision to smoke or not smoke.
210 See the Logic of Evaluation section of this chapter.
evaluating research, is sometimes a precursor to other purposes, namely the aforementioned process of selecting from amongst research proposals, which includes prioritizing (e.g., which research projects are most important?), given that the results of most research are largely unknown—which brings one to the fact that decision making often requires making predictions.

Prediction, though not a fundamental purpose of evaluation, is almost unavoidable in the evaluation of research and researchers. As Salmon (1998) points out, there are at least three—probably more—legitimate reasons for making predictions. First, predictions are made on the basis of simple curiosity about future events, without waiting for the events to transpire. Second, predictions are often made for the sake of testing a theory or hypothesis. Third, there are situations in which some practical action is required, and the choice of optimal actions depends upon predicting future occurrences. It is the third case which is of interest in the evaluation of research, particularly in regards to researchers. However, this is not the type of prediction which deals with the predictive aspect of scientific knowledge embodied in the predictive content or power of a scientific theory, for instance. It is the prediction of future performance on the basis of past performance.

While the previous sections of this chapter have outlined and discussed the historical origins of evaluation in and of research, the fundamental logic of evaluation, and the primary reasons and motives for evaluating research, the remainder of the chapter is organized around the often particularistic criteria applied to research in the form of norms and standards such as the scientific method, the ethos of science, and various notions of methodological rigor. Finally, the chapter concludes with a brief discussion of some of the demonstrable properties of good research, including its definition and a short list of merit-defining criteria.
The essential elements of a scientific method—applying mostly to experimental sciences—are iterations, recursions, and orderings of the following:

**Characterizations.** The scientific method depends upon increasingly more sophisticated characterizations of subjects of the investigation. The subjects can also be called lists of unsolved problems or the unknowns. While seeking the pertinent properties of the subjects, this may also entail definitions and observations; the observations often demand careful measurements.

**Hypotheses.** A hypothesis is a suggested description, explanation, or predication of the subject. Sometimes, but not always, they can also be formulated as existential statements, stating that some particular instance of the phenomenon being studied has some characteristic or causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic. In general scientists tend to look for theories that are ‘elegant’ or ‘beautiful.’ However, if a model is mathematically complex, it is difficult to deduce any prediction.

**Predictions.** Any useful hypothesis will enable predictions, by reasoning including deductive reasoning. It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction can also be statistical and only talk about probabilities. It is essential that the outcome be currently unknown. Only in this case does the eventuation increase the probability that the hypothesis be true. If the outcome is already known, it is called a consequence and should have already been considered while formulating the hypothesis. If the predictions are not accessible by observation or experience, the hypothesis is not yet useful for the method, and must wait for others who might come afterward, and perhaps rekindle its line of reasoning. For example, a new technology or theory might make the necessary experiments feasible.

**Experiments.** Once predictions are made, they can be tested by experiments. If test results contradict predictions, then the hypotheses

---

212 Popper (1963/2006), following others, has argued that a hypothesis must be falsifiable, and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. It must at least in principle be possible to make an observation that would show the proposition to be false, even if that observation had not yet been made.

213 Often referred to as parsimony, in reference to Ockham’s Razor, a principle attributed to the 14th century English logician and Franciscan friar William of Ockham. It forms the basis of methodological reductionism, also called the Principle of Parsimony or Law of Economy, and has become a basic perspective for those who follow the scientific method.
The essential elements of a scientific method—applying mostly to experimental sciences—are iterations, recursions, and orderings of the following:

**Characterizations.** The scientific method depends upon increasingly more sophisticated characterizations of subjects of the investigation. The subjects can also be called lists of unsolved problems or the unknowns. While seeking the pertinent properties of the subjects, this may also entail definitions and observations; the observations often demand careful measurements.

**Hypotheses.** A hypothesis is a suggested description, explanation, or predication of the subject. Sometimes, but not always, they can also be formulated as existential statements, stating that some particular instance of the phenomenon being studied has some characteristic or causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic.\(^{212}\) In general scientists tend to look for theories that are ‘elegant’ or ‘beautiful.’ However, if a model is mathematically complex, it is difficult to deduce any prediction.\(^{213}\)

**Predictions.** Any useful hypothesis will enable predictions, by reasoning including deductive reasoning. It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction can also be statistical and only talk about probabilities. It is essential that the outcome be currently unknown. Only in this case does the eventuation increase the probability that the hypothesis be true. If the outcome is already known, it is called a consequence and should have already been considered while formulating the hypothesis. If the predictions are not accessible by observation or experience, the hypothesis is not yet useful for the method, and must wait for others who might come afterward, and perhaps rekindle its line of reasoning. For example, a new technology or theory might make the necessary experiments feasible.

**Experiments.** Once predictions are made, they can be tested by experiments. If test results contradict predictions, then the hypotheses

---

\(^{212}\) Popper (1963/2000b), following others, has argued that a hypothesis must be falsifiable, and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. It must at least in principle be possible to make an observation that would show the proposition to be false, even if that observation had not yet been made.

\(^{213}\) Often referred to as parsimony, in reference to Ockham’s Razor, a principle attributed to the 14th century English logician and Franciscan friar William of Ockham. It forms the basis of methodological reductionism, also called the Principle of Parsimony or Law of Economy, and has become a basic perspective for those who follow the scientific method.
are called into question and explanations may be sought. Sometimes experiments are conducted incorrectly and are at fault. If the results confirm the predictions, then the hypotheses are considered likely to be correct but might still be wrong and are subject to further testing. Depending on the predictions, the experiments can have different shapes. It could be a classical experiment in a laboratory setting, a double-blind study or an archeological excavation. Scientists assume an attitude of openness and accountability on the part of those conducting an experiment. Detailed recordkeeping is essential to aid in recording and reporting on the experimental results, and providing evidence of the effectiveness and integrity of the procedure. They also assist in reproducing the experimental results.

A linearized, pragmatic scheme of the four essential elements of the scientific method is sometimes offered as a guideline for scientists in conducting research in the form of (Gauch, 2003):

1. Define the question
2. Gather information and resources
3. Form hypothesis
4. Perform experiment and collect data
5. Analyze data
6. Interpret data and draw conclusions that serve as a starting point for new hypotheses
7. Publish results

This schema is currently accepted as ‘standard’ scientific method. However, a number of philosophers, historians, and sociologists of science (e.g., Feyerabend, 1975) claim that it has little relation to the way science is actually practiced nor does it represent alternative notions of scientific method (Scriven, 2006b), such as Haig’s (1995) abductive explanatory inferentialism or other forms of scientific inquiry (Denzin & Lincoln, 2000;

\[214\] In the social sciences, in particular, randomization is purported to control for an infinite number of rival hypotheses without ever specifying what any of them are. However, randomized assignment never completely controls these rivals, but renders them implausible to a degree estimated by the statistical model (Campbell, 1994).
Flick, 2006; Glaser & Strauss, 1967; Guba & Lincoln, 1989; Schwandt, 2000). Ultimately, however, the evaluation of research is not about components of scientific method, it is about the user of the methods, no matter what, if any, perspective is held by the user, and:

...scientific knowledge is composed not of eternal truths but rather those of empirical findings and theories that have gained sufficient acceptance among scientists to be taken as the pertinent intellectual context for their current work. The strongest conditional guarantee that can be given, therefore, is the promise of continual research to determine how far, and in what ways, currently accepted knowledge may be wrong (Brewer & Hunter, 1989, pp. 38-39).

Ethos of Science

Partly as a response to the Nazis’ attempts to control science, Merton wrote The Normative Structure of Science in 1942, which is considered one of the first systematic attempts at identifying and establishing macro-level norms for scientific research and behavior (Langfeldt, 2002). According to Merton, there is an ‘ethos of science’ consisting of four classes of scientific imperative: (1) universalism; (2) communism; (3) disinterestedness, and (4) organized skepticism.

1. Universalism. Claims are judged, and accepted or rejected, through “preestablished impersonal criteria consonant with observations and with previously confirmed knowledge” (Merton, 1942/1973a, p. 270). That is, scientific results should be analyzed objectively and be verifiable or repeatable. The claimant’s own personal or social attributes are irrelevant to the validity of truth claims. Scientific truths should be observable or testable regardless of national, political, or religious boundaries. Merton acknowledges that although “universalism is deviously affirmed in theory and suppressed in practice”, it remains “a dominant guiding principle” (Merton, 1942/1973a, p. 273).

215 Standards are also now under development for action and emancipatory research (Pawson, 2006).
2. **Communism.** Science is a communist activity in that scientists share their work with their community for the common good, and that 'the substantive findings of science are a product of social collaboration and are assigned to the community...the scientist's claim to 'his' intellectual 'property' is limited to that of recognition and esteem..." (Merton, 1942/1973a, p. 273). Essentially, progress in science comes through cooperation and collaboration between individual scientists, and between generations of scientists.

3. **Disinterestedness.** Scientists should have no emotional or financial attachments to their work, and reward comes through recognition of scientific achievement, not through monetary gains.

4. **Organized Skepticism.** "The scientific investigator does not preserve the cleavage between the sacred and the profane, between that which requires uncritical respect and that which can be objectively analyzed" (Merton, 1942/1973a, pp. 277-278). In other words, scientists should wait until 'all the facts are in' before a judgment is made.

According to Merton (1957/1973b), the norms and values of the scientific ethos are held to be binding on the man of science and are expressed in the form of prescriptions, proscriptions, preferences, and permissions. The ethos has not been codified, but can be inferred from the moral consensus of scientists as expressed in use and wont as well as in the countless writings on scientific spirit and moral indignation directed toward contraventions of the ethos (Merton, 1957/1973b). This does not mean that the norms are not violated. The desire for recognition may lead to fraud and plagiarism, violating all sets of norms (Frankel, 2005, June). Nevertheless, the idea seems to be that the ethos is either violated or obeyed, and is non-negotiable (Langfeldt, 2002). That is, violations either occur or they do not. Therefore, this is a simple matter of asymmetry (Scriven, 2006d).

Of course, the Mertonian school as it came to be known, provoked a prolonged and heated dialogue among historians, philosophers and sociologist of science, with contending groups campaigning against them (Zuckerman, 1988). One example
frequently referred to is the apparent omission of parsimony,\textsuperscript{216} which has been countered by statements such as: “divine providence is a parsimonious explanation of just about anything” (Collins, 1982, p. 299), therefore, it is asserted by some that parsimony has no basis as a scientific norm or standard.

As Langfeldt (2002) observed in her study of expert panel decision processes,\textsuperscript{217} however, the dispute is not whether or not Merton’s ethos, are universally accepted scientific standards and norms, but, whether violations of them occur, and they are easy to find, for example:

...Soman’s fabrication of data and his retraction in 1979 of twelve papers, the majority published in collaboration with the holder of an endowed chair at Yale University Medical School...the biologist, Alsabti’s rash of plagiarized papers, which came to light in 1980...Spector’s unreplicable explanations of virus as a unified cause of cancer, aired in 1981...the unfolding in the same year of Darsee’s fabrication of data which resulted in the publication of over one hundred papers while at Emory and at Harvard...the 1986 announcement of a University of California-San Diego committee that nearly half of the 147 articles of a rising radiologist, Slutsky, were found to be “fraudulent” or “questionable”...the cases of Glueck’s misrepresentation of data on cholesterol and heart disease...Bruenig’s articles, based upon nonexistent experiments of psychotropic drugs to control behavior of the mentally retarded within institutions, both of which came to light in 1987...the seemingly endless affair of the disputed paper in \textit{Cell} by Imanishi-Kari with Nobel laureate David Baltimore as one of five other coauthors... (Fox, 1994, p. 298).

Jan Hendrik Schöen...from 1998 to 2001, published an average of one research paper every eight days—alone and with co-authors—of which 17 of those papers had come out in \textit{Nature} and \textit{Science}. Due to allegations of scientific misconduct a committee was set up in 2002 in order to investigate possible scientific fraud. In the final report by the commission evidence of Schöen’s misconduct was shown in at least 16 out of 24 allegations (Dinges, 2006, p. 12).

\textsuperscript{216} Parsimony here is in reference to Ockham’s Razor, a principle attributed to the 14\textsuperscript{th} century English logician and Franciscan friar William of Ockham. It forms the basis of methodological reductionism, also called the Principle of Parsimony or Law of Economy, and has become a basic perspective for those who follow the scientific method.

\textsuperscript{217} Langfeldt’s (2002) study dealt with the constraints on, processes in, and biases of expert panels evaluating research quality and research priorities.
Therefore, “the real test of the significance [of norms and standards]...” it is argued, is “[do]...violations...particularism, non-communication of results, personal bias or ‘dogmatism’...sanction or entail moral indignation among colleagues...if detected?”

and:

...the answer depends on how categorically the ethos of science is interpreted. If we take universalism [italics added] to be incompatible with all kinds of ‘old boys networks’ and institutional and personal loyalties, communism [italics added] to be a norm not only for academic research, but also for industrial research, organised skepticism [italics added] to forbid the kind of dogmatism a scientific paradigm entails, and disinterestedness [italics added] to be incompatible with the kind of personal bias resulting from having personal and institutional loyalties...then we do not even need empirical studies to answer the question. With such an interpretation there are obvious situations in which one would be expected to be (and rewarded for being) ‘particularistic’, ‘dogmatic’, ‘personally biased’ or to not communicate results. On the other hand, if we adopt a ‘soft’ interpretation, and say that the only violations of the ethos are those including behavior that are clearly understood as fraud or misconduct, we come close to a tautological argument, saying that all violations of the ethos entail more indignation because everything defined as fraud or misconduct entails moral indignation...the fact that some behaviour is defined as fraud or misconduct and sanctioned, is evidence enough to claim that the scientific community has some norms (Langfeldt, 2002, pp. 52-53).

Furthermore, Merton himself stressed that “the social institution of science...incorporates potentially incompatible values” (1963/1973c, p. 383). On the one hand, there is a “value set upon originality which leads scientists to want their priority recognized” (Merton, 1963/1973c, p. 383). On the other, there is the norm of “selfless dedication to the advancement of knowledge for its own sake” (Merton, 1963/1973c, p. 399).

Although the debates in the history, philosophy, and sociology of science continue, in the decades that followed The Normative Structure of Science (Merton, 1942/1973a), norms and standards of scientific practice at a meso-level (Liljenström & Svedin, 2005) have generally been in reference to methodological axioms. When it comes
to applying notions of quality in practice, the debate about what constitutes methodological quality has, to some extent, been fueled by the widespread influence of the hierarchy of evidence used in health care (Boaz & Ashby, 2003).

Methodological Rigor

Methodologists are the preachers of science. Armed with canons of correct procedure, they have the power to castigate and exhort. They can instruct us to have clearly defined objectives and explicit frames of reference, to base our studies on good theories...the process of science does not work from rules to practice but from attempt to attempt...one good piece of research influences research practice more than many methodology textbooks (Przeworski, 1987, p. 31).

Rigor in research is normally conceived of as the means by which integrity and competence are confirmed (Tobin & Begley, 2004). That is, a way of demonstrating the legitimacy or soundness of the research process. Without rigor, it is argued, there is a danger that research may become fictional journalism and therefore worthless as contributing to knowledge (Morse, Barrett, Mayan, Olson, & Spiers, 2002).

A substantial proportion of the scientific community associates research quality with methodological rigor (Farrington, 2003), though it only constitutes a small segment of the scientific method. However, since its introduction by Donald Campbell and Julian Stanley in the early 1960s, followed by its successors in 1979 (Cook & Campbell) and 2002 (Shadish, Cook, & Campbell), hierarchies of evidence, as they have come to be known in some circles, often form the basis by which research quality is judged.

218 Rigor literally means 'the quality of being unyielding or inflexible' or 'strict precision.'
219 Parts of this section appeared in the Journal of MultiDisciplinary Evaluation (Coryn, 2007b).
220 Many undergraduate and graduate programs now include courses intended to make them good consumers and evaluators of research using textbooks, such as Evaluating Research Articles (Girden, 2001), which is intended to train students in evaluating the soundness of research designs and appropriateness of statistical analyses in the published research literature.
221 This does not imply that standards of methodological rigor did not exist prior to Campbell and Stanley's Experimental and Quasi-Experimental Designs for Research (1963), only that these have become the standards by which much research has been upheld in the last four decades.
Foremost amongst users of these types of hierarchies are the health sciences (Boaz & Ashby, 2003; Grayson, 2002). The hierarchy of evidence often employed to judge methodological rigor, or soundness, especially of quantitative research is (in descending order from highest to lowest quality, or rigor): 222

1. Systematic reviews and meta-analyses 223
2. Well-designed randomized controlled trials
3. Well-designed trials without controls (e.g., single-group pre-post, time series or matched case-controlled studies)
4. Well-designed non-experimental studies from more than one center
5. Opinion of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees

These hierarchies are grounded in the quantitative tradition and usually consist of the criteria of validity, reliability, and objectivity. Combined, these criteria have almost reached the status of a ‘holy trinity’ (Kvale, 1995; Tobin & Begley, 2004). 224 Yet, the exact nature of validity has eluded adequate, and agreed upon, characterization since there exists no single or common explanation of the term. Common definitions include “[research is valid]...if it represents accurately those features of the phenomena that it is intended to describe, explain or theorise” (Hammersley, 1987, p. 69) and “the truth of, correctness of, or degree of support for an inference” (Shadish, Cook, & Campbell, 2002, p. 513). 225 The traditional view of reliability, on the other hand, is premised on

---

222 There are many variations of hierarchies related to methodological rigor. The one presented here is primarily applied to address the effectiveness of clinical interventions.
223 The placing of meta-analysis and systematic reviews on the hierarchal ladder, let alone at the top, is interesting given that there are incredible variations in the quality of the two, and they are not necessarily viewed as research designs per se, but as methods to review literature.
224 Other versions of the trinity often include generalization rather than objectivity as a criterion for judging the soundness, rigor, or quality of research (Kvale, 1995). However, internal and external validity are usually considered the essential elements of generalization, especially the latter. For instance, statistical sampling is often the dominant basis for generalizing (Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2004; Kish, 1965).
225 In fact, there is presently no ‘index’ of validity.
assumptions of replicability, repeatability, or consistency (Golafshani, 2003; Shadish, Cook, & Campbell, 2002). However, reliability is often portrayed as the extent to which a measurement or observation yields the same answer or results however and whenever it is carried out (i.e., time and place independent). As a pair, validity and reliability are sometimes described as two complementary aspects of objectivity (Tobin & Begley, 2004) to the extent that objectivity is normally understood a being free from bias (e.g., cognitive, cultural, sampling) or distortion (Trochim, 2002).

Debate around the relevance and use of this and other versions of the trinity as standards of research quality, or methodological rigor, has persisted for more than 20 years (Flick, 2006; Guba & Lincoln, 1981; Morse, Barrett, Mayan, Olson, & Spiers, 2002; Sandelowski, 1986; Tobin & Begley, 2004; Winter, 2000; Yin, 1994), and Guba (1981) warned that these criteria are “primitive” (p. 90), and should be applied only as guides rather than orthodoxy. By far, the most energy has been devoted to developing standards for assessing the quality of randomized controlled trials (RCTs), which are usually:

1. Was the assignment to the treatment groups really random?
2. Was the treatment allocation concealed?
3. Were the groups similar at baseline in terms of prognostic factors?

In order to estimate reliability, researchers often apply various theoretical or statistical assumptions such as true score theory (Spearman, 1907, 1913) or generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). However, reliability in terms of design should not be confused with reliability from the test theory notions. The reliability criterion of methodological rigor is more about replicability, repeatability, or consistency, whereas in test theory it is more a statement about common variance shared.

Unlike geology, volcanology, and similar sciences which are time and place dependent.

However, in science, the ideal of objectivity is generally considered to come about as a result of strict observance of the scientific method. Objectivity in science is intimately related to the aim of reproducibility. Methodological aspects can be roughly distinguished as objectivity in measurement and objectivity in experimentation and interpretation. As such, it is only tangentially related to the concept of objectivity in philosophy, and closer to, for example, objectivity in journalism. Another methodological aspect is the avoidance of bias, which can involve cognitive bias and cultural bias, but also sampling bias. Methods for avoiding or overcoming such bias include random sampling and double-blind trials.

Randomization purports to control an infinite number of rival hypotheses without ever specifying what any of them are, and when RCTs are compromised there is a tendency for the results of systematic reviews and meta-analyses to be distorted (Juni, Altman, & Egger, 2002).
4. Were the eligibility criteria specified?

5. Were outcome assessors blinded to the treatment allocation?  

6. Was the care provider blinded?

7. Was the patient blinded?

8. Were the point estimates and measures of variability presented for the primary outcome measure?

9. Did the analyses include an intention to treat analysis?

In part, these standards have been the result of social scientists trying to replicate the natural sciences, such as physics and chemistry, as the ideal embodiments of scientific inquiry (Scriven, 2006e; Shadish & Fuller, 1994). Nowhere is this dogmatic view of methodological rigor more apparent than the United States Department of Education’s Institute for Educational Science (IES) which allocates and controls a budget of nearly $500 million and is now only funding RCTs (Donaldson & Christie, 2005; Scriven, 2006, September).

Criteria for assessing the quality of qualitative research have also emerged in order to address the ‘fitness for purpose’ of research, although, their application often involves some redefinition of the terms:  

The usual canons of good science have value but require redefinition to fit the realities of qualitative research and the complexities of the social phenomena that we seek to understand (Strauss & Corbin, 1998, p. 266).

Like the quantitative tradition, most criteria for assessing qualitative research have come from the field of health studies (Boulton & Fitzpatrick, 1994; Lincoln, 1992; Lincoln & Guba, 1990; Mays & Pope, 1995). One of the most widely applied lists comes

---

230 Blinding is sometimes referred to as masking (Shadish, Cook, & Campbell, 2002).
231 Fitness for purpose is the notion that in assessing the quality of research that methodological rigor is but one aspect of quality; the other aspect is the relevance of the research for policy or practice (Boaz & Ashby, 2003; Patton, 2002).
from the Medical Sociology Group (1996), which includes the following criteria for assessing the soundness of qualitative research:

1. Are the research methods appropriate to the question being asked?
2. Is there a clear connection to an existing body of knowledge?
3. Are the criteria for/approach to sample selection, data collection and analysis clear and systematically applied?
4. Is the relationship between the researchers and researched considered, and have the latter been fully informed?
5. Is sufficient consideration given to how findings are derived from the data and how the validity of the findings were tested?
6. Has evidence for and against the researcher's interpretation been considered?
7. Is the context for the research adequately described and accounted for?
8. Are findings systematically reported and is sufficient original evidence reported to justify a relationship between evidence and conclusions?
9. Are researchers clear about their own position in relation to the research topic?

From this and similar lists, the qualitative tradition ultimately developed its own holy trinity, which includes the criterion of trustworthiness (including subcriteria of credibility and transferability), the criterion of dependability, and the criterion of confirmability (Flick, 2006; Golafshani, 2003; Guba & Lincoln, 1989; Trochim, 2002).\footnote{Flick (2006), for example, discusses qualitative validity in terms of “whether researchers see what they think they see” (p. 371). Moreover, he and others (Kirk & Miller, 1986) argue that three types of error may occur as regards qualitative validity: seeing a relationship, a principle, and so on when they are not correct (Type I error); to reject them when they are correct (Type II error); and asking the wrong questions (Type III error). Type III error occurs when the right answer to the wrong question is observed (Raiffa, 1968).}

Trustworthiness is the extent to which the results of qualitative research are credible...
or believable from the perspective of the participant in the research (i.e., credibility) and the degree to which results can be generalized or transferred to other contexts or settings (i.e., transferability). The dependability criterion emphasizes the need for the researcher to account for the ever-changing context within which research occurs as most qualitative researchers tend to assume that each researcher brings a unique perspective to the study. Confirmability refers to the degree to which the results can be confirmed or corroborated by others (Trochim, 2002).

To the extent that the generally accepted quantitative and qualitative criteria of methodological soundness, or quality, differ is subject to dispute, but the similarities are strong enough that it can be reasonably inferred that they are in fact comparable. For instance, the qualitative subcriteria of the trustworthiness criterion (i.e., credibility and transferability) are simply parallels of the quantitative concepts of internal and external validity, where credibility is synonymous with internal validity and transferability is congruent to external validity. The dependability criterion, on the other hand, is analogous in most respects to the quantitative criterion of reliability and confirmability is in essence the quantitative standard of objectivity.

Although the quantitative and qualitative trinities represent ‘aspects’ of good research, in and of themselves, they are not sufficient to judge the merits of an instance of research, and certainly not its worth or significance. In any case, rigor in terms of integrity, competence, legitimacy, or soundness is one of sufficiency. The correct standard, or basis, in science or outside it, for such conclusions is that they can be demonstrated or established beyond reasonable doubt (Scriven, 2006e).

Beyond reasonable doubt, “the standard of evidence that the courts require in felony trials, is far stronger than ‘the balance of evidence’ the standard they use, and sharply distinguish, for misdemeanor trials. People sometimes think that RCTs are the paradigm design because they meet some higher standard than beyond reasonable doubt, perhaps ‘beyond the practical possibility of error.’ But they are far from that standard, which is not even met by proofs in deductive logic and mathematics” (Scriven, 2006e, p. 4)
Properties of Good Research

In nearly all cases, the fundamental task required for evaluating researchers and their research is identifying and validating the properties or characteristics which define the scale of merit for a 'good' researcher or 'good' research. Is there something that is irrefutable as the basis for evaluating research? Is there something that must be accepted as the nature or essence of good research? Are there in fact identifiable, confirmable criteria for good research?

Definitions of Research

One of the most serious problems encountered in evaluating research, or researchers, is the dispute about what is and is not research (e.g., Alcorn, Cardno, Fairburn-Dunlop, Jones, O'Brien, Bishop, Crooks, Hattie, Kane, & Stevenson, 2006; Boyer, 1990; COSEUP, 2001; Glassick, Huber, & Maeroff, 1997; New Zealand Qualifications Authority, 1998; OECD, 2002b; Tertiary Education Commission, 2004a). That is, how is research defined? Or, can it be satisfactorily defined? Of course, there are numerous competing definitions of the term and little apparent consensus (Calvert & Martin, 2001). However, common usage holds that research is a truth-seeking activity which contributes to knowledge, aimed at describing, predicting, or explaining the world, conducted and governed by those with a high level of proficiency or expertise (Coryn, 2006d). Some (e.g., Boyer, 1990; Calvert & Martin, 2001; COSEUP, 2001; Glassick, Huber, & Maeroff, 1997; OECD, 2002b) would object to this definition, however, and claim that research should be distinguished and separately defined as basic or applied (i.e., by including a research purpose or objective). Since both basic and applied research are

---

235 Parts of this section appeared in the Journal of MultiDisciplinary Evaluation (Coryn, 2006d).
236 The second part of this definition is that research is also a particular instance or piece of research (Coryn, 2006d).
contributions to knowledge, the need for the distinction is questionable and the proposed
definition adequately captures the essential nature of research, is neither too narrow nor
too broad, and simply avoids complications.237

This definition consists of three distinct and equally significant parts. First, research is truth seeking. Truth seeking is the search, or investigation, of or for a body of real things, events, or facts. Second, research describes or explains (the “what’s so”). To describe involves representing or giving an account of. To explain is to give the reason for or cause of. Combined, or separately, these two parts result in a contribution to knowledge. Third, research is conducted and governed by those who have the requisite proficiency or expertise. To be proficient or to be an expert means that one is well advanced in a branch of knowledge derived from training or experience. There might seem to be some circularity here, however, because one is proficient or an expert does not imply that one contributes to knowledge. It only implies that the latter is necessary, but not sufficient, for doing research. Therefore, it may also be claimed that the task or duty of researchers is truth seeking, aimed at describing or explaining, conducted at a high level of proficiency or expertise, which results in a contribution to knowledge.

By contrast, the amalgamation of several typical dictionary definitions (e.g., Merriam-Webster’s Collegiate Dictionary, 2003; Oxford American Dictionary [McKean], 2005; Oxford English Dictionary [Soanes & Stevenson], 2004) produces a composite definition of ‘studious inquiry or examination, or investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws.’ While the composite of dictionary definitions lists many of the critical qualities that express the

237 Scriven (2007) recently defined science as ‘interested in describing, predicting, or explaining the world and the important events in it—not all events, but just those that contribute significantly to our useful knowledge and understanding of the world.’
essential nature of research (e.g., studious inquiry or examination), they also omit one of the key features (i.e., conducted and governed by those with a high level of proficiency or expertise).

It has also been asserted that most definitions of research fail to account for, and exclude, much of the work done in the arts and humanities. In the creative arts alone, it has been argued that these definitions are insensitive, and create anomalies as to what constitutes research (Research Evaluation and Policy Project, 2005; Strand, 1998). For instance, a painter’s paper about their own exhibition might qualify as research, whereas the painting does not under the definitions given above. A critical paper on a musical composition might succeed in meeting the definition from standard usage or the typical dictionary definition for being considered research, while the performance of the composition, and even the composition itself, would not (Strand, 1998). Nonetheless, research is a cognitive activity, not an aesthetic one, and, in many instances, the creative arts are not clearly cases of research. By contrast, the accumulation of related case histories in law is a research endeavor—and the discipline argues strongly that it is—in that it constitutes both truth seeking as well as description and it is certainly a contribution to knowledge.238

Similar problems arise in definitions that define research in terms of research typologies (i.e., basic research, applied research, experimental development).239 For instance, one of the most widely applied definitions of research is from the OECD’s Frascati Manual (2002), in which research is classified into three categories:240

238 It could also be reasonably inferred that many cases of evaluation meet the definitional criteria for research, particularly ascriptive evaluation.
239 While not entirely irrelevant, for example, in assessing the needs for research (see the Values and Valuing subsection of this chapter), the typological distinction is not always necessary for determining whether research is good, valuable, or important.
240 The Frascati Manual was first published in 1963, and is now in its 6th edition. The manual primarily deals with measuring the resources devoted to R&D.
Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.

The term R&D covers three activities: basic research, applied research and experimental development. Basic research is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particular application or use in view. Applied research is also original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific practical aim or objective. Experimental development is systematic work, drawing on existing knowledge gained from research and/or practical experience, which is directed to producing new materials, products or devices, to installing new processes, systems and services, or to improving substantially those already produced or installed (OECD, 2002b, p. 30).

These typologies occur in nearly all definitions of research, except those from standard usage and to some extent the typical dictionary definitions, and their usefulness for evaluating research is uncertain. Research does not require this disaggregation in order to meet the definitional requirement of a description, explanation, or contribution to knowledge; it is either a contribution to knowledge or it is not (Scriven, 2006d), and the aims and objectives (e.g., basic versus applied), in most cases, are irrelevant.

Other definitions, for instance, the Carnegie Foundation’s definition of research, include teaching as part of its definition. The Carnegie definition, from Scholarship Assessed: Evaluation of the Professoriate (Glassick, Huber, & Maeroff, 1997), consists of four parts, which are: 241, 242

1. Discovery. Contributes not only to the stock of human knowledge but also to the intellectual climate of an institution

241 The rationale for this definition is to “ensure that scholarly work in areas both within and outside discovery can be appropriately recognized and rewarded” (University Grants Committee, 2004, p. 22) in Hong Kong’s Research Assessment Exercise (RAE).

242 This definition is conceptually similar to that from an earlier Carnegie study by Boyer (1990), which appeared in Scholarship Reconsidered: Priorities of the Professorate.
2. Integration. Work that seeks to interpret, draw together and bring new insights to bear on original research.


4. Teaching. Transforms and extends knowledge while transmitting an intelligible account of knowledge to the learners.

The major conceptual problem with this definition of research is that only the first part approximates the standard definitions of research, however, the second, third, and fourth parts do not. The second and third parts are not conditional requirements for research (i.e., description, explanation, and knowledge generation), they are aims (i.e., integrating) and uses (i.e., application). The fourth part is an attempt to integrate the teaching and research functions of higher education faculty. Definitions such as this are seductive in that they suggest that if one is a good or productive researcher then one might also be a good or effective teacher (Marsh & Hattie, 2002), and that transferring knowledge as opposed to generating it is a form of research, which it is not. Teaching and research are two independent tasks or duties; that is, “different enterprises” (Hattie & Marsh, 1996, p. 513).  

Unlike other characterizations of research, the United States Committee on Science, Engineering, and Public Policy (COSEUP)—a joint committee consisting of the National Academy of Sciences (NAS), National Academy of Engineering (NAE), and Institute of Medicine (IOM)—has portrayed research in such a way as to make it possible for federally-funded agencies to meet the Government Performance and Results Act (GPRA) of 1993 reporting requirements.  

243 Hattie and Marsh (1996), in their meta-analysis of 58 studies, found only a small positive relationship (weighted $r = .06$) between research and teaching.

244 Many nations include not only teaching in their evaluations of federally-funded research, but also in many cases the number of degrees awarded, neither of which are indicators of good research or good researchers; although, the latter might be an indicator of a good research program.

245 GPRA requires agencies to produce three documents: a strategic plan, a performance plan, and a performance report (COSEUP, 2001).
...[research is] a search for the unknown whose outcomes are virtually unlimited, research defies exact definition. Intellectually, it is apparent that the performance of research takes place across a continuum of thought and action, from the abstract reasoning of a single individual to a multi-billion dollar program of technological complexity, such as a mission to Mars.

...[to satisfy administrative and intellectual needs] it has often been convenient to separate "basic" research from "applied" research...

...[in managing and funding research] it is important to understand the open-ended possibilities of any research activity, no matter how it is categorized, and to encourage the freedom of inquiry that leads beyond what is already known (COSEUP, 2001, p. 8).

If research is an indefinable nebulous concept that defies accurate characterization, how is it that nearly $125 billion of the United States' budget was allocated to it for 2006 (U.S. Office of Management and Budget, 2005)?246 Nevertheless, implicit in COSEUP's (2001) depiction is that, by its very nature, research requires some form of intellectual investigation (i.e., truth seeking) which contributes to understanding.

While the definitional debate appears trivial, or pedantic, the fact remains that the fundamental intent of an accurate definition is not only to describe the essential nature of something, but also to begin the process of identifying merit-defining values and criteria, many of which are drawn from these definitional properties and true characteristics.247

The Search for Criteria

In 1963, Weinberg distinguished between two categories of criteria for evaluating scientific research: internal and external. Internal criteria are those that arose "from within the science itself, and are basically criteria of efficiency" (Weinberg, 1989, p. 4).

---

246 This includes the defense and non-defense research budget allocation, which represents 4.9% of the total United States federal budget for 2006 (U.S. Office of Management and Budget, 2005).

247 See the Values and Valuing and Criteria versus Indicators subsections of this chapter.
That is, "how well proposed research would be performed" (Weinberg, 1989, p. 4). External criteria were "criteria of utility; that is, the degree to which the given research, if successful, is useful outside the field itself" (Weinberg, 1989, p. 4), and were differentiated into three subcriteria: technological merit, social merit, and scientific merit. These criteria were prompted by "a drive to evaluate proposals for research, not research already done" and for determining "whether or not to go to the moon, and how much should go for the support of high energy physics" (Weinberg, 1989, p. 4-5). Ultimately, Weinberg argued that these criteria:

...involve the concept of 'value' within science. In saying this I distinguish between the 'value' and 'truth' of a scientific finding. Every valid science finding in some sense reveals an aspect of truth...in the allocation of resources—that is, the administration of science—truth alone is insufficient. Not all scientific activities that meet the strictest criteria of scientific truth can be supported; one must judge which activities are most valuable. Thus the notion of value within science is at the heart of the practice of science. So to speak, the philosophy of scientific administration is axiological, and the philosophy of scientific practice is epistemological (Weinberg, 1989, p. 5).

In principle, the evaluation of research hardly needs elaborate philosophical underpinnings as there is normally, but not always, a consensus around what is truly important research. However, Weinberg's (1963, 1989) notion of internal and external criteria are limited and they fail to identify the complete array of relevant values, including social values, and does little in the way of making the evaluation of researchers and their research explicit, and therefore they are incomplete; still, they were a major breakthrough.

Once again, evaluation in research can be understood as implicit evaluation conducted by scientists and researchers, as part of intra-scientific practice, such as evaluation of the quality of evidence, research designs, instruments, and so forth. Thus, the criteria normally applied for this type of evaluation are implicit criteria or values, which
are often, but not always, subjective. By comparison, evaluation of research may be understood as *explicit evaluation* of research or researcher quality, value, or significance, which is considered an extra-scientific as well as intra-scientific practice. Therefore, *explicit criteria* or *values* need to be identified.

**Implicit Criteria**

Historically, science has been a self-governing, self-evaluating endeavor, generally based on a constitutive perspective of what represents good research (Langfeldt, 2002). The constitutive perspective claims that there are certain characteristics that represent good research as such, yet restrict what may be meant by good research, mostly on the basis of implicit criteria or values and differing ontological views as to what is understood as quality dimensions of research. Implicit criteria or values are those such as the criteria suggested by Kuhn (1977/1998) to be applied for evaluating the adequacy of a theory (i.e., accuracy, consistency, scope, simplicity, and fruitfulness). These criteria are those by which researchers operate and practice their day-to-day work. They are the criteria by which researchers evaluate their own work and the work of their predecessors, peers, and colleagues; that is, the evaluation of the previous literature, hypotheses, theories, and research designs, for example.

Yet, there is no general agreement on the specific constitutive attributes of good research that refer to criteria or values in an objective or systematic way. More often than

---

248 The constitutive perspective claims that there are standards of good research unrelated to what evaluators might define as good research. Moreover, this perspective asserts that research quality may be constituted in terms of idealism and realism, as well as pragmatism. Realism says that there is 'something' characteristic of good research, whereas idealism says that good research is socially and culturally constructed. The perspective is also related to nominalism, which implies that research quality is an empty concept meaning that good research, researchers, programs, or institutions have nothing in common except "we say that they are good" (Langfeldt, 2002, p. 47).

249 Implicit values are those most often applied by reviewers of manuscripts submitted for publication or proposals for funding.
not, the constitutive perspective of good research judges a theory as good because it has explanatory power or a research question as good because it is researchable. For instance, this assumes that a question needs specific attributes, for example, researchability, fruitfulness, stringency, and originality, to be a good research question and implying that a question lacking some or all of these characteristics cannot be a good research question (Langfeldt, 2002).

These values normally turn out to be arbitrary expressions of preference; for example, that only quantitative research is valuable and qualitative research is worthless. Thus, these implicit, constitutive criteria and values are frequently paradigm-specific, grounded in a specific philosophical or epistemological perspective (e.g., that causation can only be demonstrated by RCTs),


\[250\] governed by rigid methodologism (Salmon, 2003),

\[251\] or determined by disciplinary conventions, among others.\[252\] This is not to say that constitutive values are not intrinsically evaluative, they are. However, they tend to focus too narrowly on only very limited aspects of research quality or are guided by how one understands the ontological status of research quality. Like methodological rigor, implicit values are not sufficient in themselves to judge the merits of an instance of research, and surely not its worth or significance.

Explicit Criteria

As discussed elsewhere in this chapter, one of the key tasks of systematic, objective, and explicit evaluation is the identification of the relevant and demonstrable values premises by which an evaluation of an evaluand or evaluée is upheld.\[253\] Given

\[250\] Similar in most respects to the ‘paradigm wars’ (Datta, 1994; Salmon, 2003) in psychology, evaluation, and other disciplines.

\[251\] That is, research is good if it was conducted according to certain methods.

\[252\] These types of values are not irrelevant, and in some cases they are definitive (e.g., the ‘quality’ of data or the ‘randomness’ of assignment).

\[253\] See the Values and Valuing and Criteria versus Indicators subsections of this chapter.
that the operational definition of research as applied here is that research is a truth-seeking activity which contributes to knowledge, aimed at describing, predicting, or explaining the world, conducted and governed by those with a high level of proficiency or expertise (Coryn, 2006d), many values, which can be applied across nearly all domains of research, can be inferred from this definition. To a lesser extent, values and criteria can also be drawn from the norms and standards of research (i.e., scientific method, ethos of science, and methodological rigor) presented previously, although these tend toward the implicit, constitutive type of values.²⁵⁴

Ergo, the definitional characteristics, norms and standards of research, and a number of implicit criteria, demonstrate that there are in fact relevant and confirmable values or merit-defining criteria which can be applied to many instances or pieces of research, or to a researcher, or R&D center, or in some cases a research proposal.²⁵⁵

These include, for example:

1. **Originality/novelty.** Generally, originality and novelty represent independence of thought and newness.²⁵⁶ When and if a breakthrough comes, “more intellectual credit is given to those who discover and interpret the unknown than those who confirm the work of others” (Morgan, 1985, p. 8).

2. **Significance/importance.** Significant or important research has the quality of having meaning and value; including scientific, intellectual, economic, and social (Hansson, 2002).

3. **Relevance.** Relation to the matter at hand, practical and especially social applicability, also fit for the particular purpose.

4. **Fecundity.** Intellectually productive or inventive. In the philosophy of science, it also refers to the ability to open new lines of inquiry.

---

²⁵⁴ This does not make them arbitrary, particularly if making contextually evaluative claims (e.g., research conducted within-paradigm).

²⁵⁵ A more complete research proposal evaluation checklist is currently under development (Coryn, 2007c).

²⁵⁶ Kuhn (1962/1996) and Lakatos (1970) argued that risky ‘paradigm breaking’ science, or science that goes beyond the sphere of the dominant ‘research program,’ is likely to have the greatest impact.
5. **Uniformity.** In reference to constant high quality. Uniformity is a criterion for evaluating a researcher or an R& D center, but not an item or particular instance of research. It is also useful, in many cases, for proposal evaluation.

6. **Validity.** High quality research is logically correct and credible. Or, the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions (Messick, 1989).

7. **Replicability.** The extent to which research is capable of replication or has the quality of being repeatable.257

8. **Ethicality.** Ethically-performed research conforms to a rule or habit of conduct with regard to right and wrong or a body of such rules and habits; the moral quality of a course of action.

Though not themselves sufficient to determine the quality, worth, or importance research, they are, however, necessary.258 Combined with all of the relevant and demonstrable values (e.g. needs, professional standards, resource economy), however, this list provides a nearly comprehensive inventory of the characteristics and properties of good, valuable, and important research. Nonetheless, dimensions of good research or a good researcher often differ by discipline, by method, by age of development, and by culture (i.e., contextually dependent).

---

257 Some assert that “Replicability of empirical findings is the core requirement of science. All else is embellishment” (Silver, 2007, January 11). However, “This is a common slogan but completely wrong. The Indonesian tsunami is a one-time event, and it caused plenty of specific damage, and no one in their right mind doubts it occurred, or that it caused the damage it caused, but it will never be replicated. The replicability criterion is just a hangover from the idea that basic physics and chemistry, which are mainly interested in general laws, are the only sciences. The formation of the continents on this planet wasn’t even observed, let alone replicable, but we are very knowledgeable about it…Do we want to argue that the Holocaust has to be replicable in order to show it happened? Not in the least; the evidence is decisive and the event in its particular detail, is not replicable. In general, history is factual and causal but not replicable, even in principle, though it is observable in principle, like the geological events above. People also like to say, with Popper, that refutability is the sole criterion that matters. Equally false; plenty of scientific laws, for example, the general gas laws, are about ideal entities, part of whose definition is that they make these laws true. They are not falsifiable, because they’re definitionally true. Or…they like the idea that elite agreement is the key test; no, it would include every fallacy like the flat earth, the geocentric universe…It’s the local substitute for truth, but it’s not the heart of truth, because it’s evanescent and truth isn’t” (Scriven, 2007, January 11).

258 Lester (1996), for instance, proposed the following criteria for determining the merits of research: worthwhileness; coherence; competence; openness; ethics; and credibility.
Conclusions

Evaluation has a long past, but a short history. It suffuses nearly everything and as a fundamental attribute of the human condition, is perhaps the single oldest, most important, and sophisticated cognitive process involved in reasoning and logic. As a systematic process for determining the merits, worth, or significance of things, however, evaluation is only about 4,000 years old, but in any case still very old. While evaluation in research is itself also a very old practice, which can be traced back to the ancient Greeks, it has primarily been a self-governing, implicit endeavor.

Consequently, the evaluation of research has a long way to go before it can be truly considered explicit, objective, or systematic. Be that as it may, the beginning of the journey was chronicled in this chapter and the points that were raised set the stage for the next. Large parts of scholarly research are publicly-funded and in most parts of the world government funding for research is tight, and is only getting tighter, due to, among other things, current funding priorities and large budget deficits. Almost universally, governments want answers to the questions "now what?" "how much?" "what if?" and "to whom?" They also want to discriminate, or sort, the good from the bad, the worthwhile from the worthless, and the important from the trivial. However, as it turns out, in most nations there are weaknesses in their understanding and application of the logic of evaluation.

As noted in the Preface, this dissertation sets forth to: (a) review the increasingly-large literature concerning the principles and procedures used to evaluate researchers and their research, as well as their underlying ideologies; (b) explain their shortcomings; (c) make clear why there are sufficient reasons to justify changing them; (d) demonstrate that research can be evaluated systematically and objectively; and (e) propose cogent alternatives and/or improvements to existing principles and procedures.
Ultimately, this dissertation is about the next step upward in the long process of making the implicit explicit; that is, the search for standards of merit to be applied for the evaluation of researchers and their research.
CHAPTER II

INTERNATIONAL RESEARCH EVALUATION MODELS

Large parts of scholarly research are publicly-funded and in most parts of the world government funding for research is highly contested. Under demands for greater accountability, due to constraints of diminished funding, and in the pursuit of general quality improvements, many countries have initiated systems for evaluating publicly-funded research at the national level.\textsuperscript{259} The evaluation of publicly-funded research now has a substantial tradition (Cozzens & Turpin, 2000), particularly in the European countries, dating to the early 1970s.

As discussed in Chapter I, some of the earliest efforts were undertaken in the Nordic countries, of which Sweden was the first country to carry out systematic evaluations of its research in the 1970s, followed in the mid-1980s by Finland, Norway, and Denmark (Luukkonen, 2002). Although there are vast differences in the way governments fund research around the world (Campbell, 2002; Campbell & Felderer, 1997; Cozzens & Turpin, 2000; Laudel, 2006), and a diversity of approaches to evaluating publicly-funded research (ab Iorwerth, 2005; Geuna & Martin, 2001, 2003; OECD, 1987, 1997, 2003; von Tunzelmann & Mbula, 2003), almost all now share the common purpose of relating funding to performance (ab Iorwerth, 2005; Campbell, 2002; COSEPUP, 1999, 2001; Cozzens & Turpin, 2000; Geuna & Martin, 2003;

\textsuperscript{259} Parts of this chapter were presented at the meetings of the European Evaluation Society/United Kingdom Evaluation Society (Coryn, 2006, October) and American Evaluation Association (Coryn, 2006, November).
Luukkonen, 2002; OECD 1987, 1997; RAE 2008, 2005; Scriven, 2006f). For example, as the OECD noted in its 1997 report *The Evaluation of Scientific Research*:

...research evaluation has emerged as a “rapid growth industry” ...[and]...there is an increasing emphasis on accountability, as well as on the effectiveness and efficiency of government-supported research...governments need such evaluations for different purposes: optimizing their research allocations at a time of budget stringencies; re-orienting their research support; rationalizing or downsizing research organizations; augmenting research productivity. To this end, governments have developed or stimulated research evaluation activities in an attempt to get “more value for the money” they spend on research support (OECD, 1997, p. 5).

This chapter begins by presenting a summary of the research landscape in 16 countries, as well as brief historical overviews, detailed accounts of the research structure and context, and the methods employed in those countries to evaluate publicly-funded research.

These nations are: 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 (see Figure 5).

---

260 The methods employed in each of the 16 countries are referred to as *Evaluation Model* throughout the remainder of this chapter. The term model is used loosely and is in reference to a collection of policies, methods, and approaches.

261 Countries included were selected on the basis of two criteria: (i) adequacy of information and (ii) availability of information in the English language.

262 Many of these descriptions were verified by independent experts.

263 The descriptions presented in this chapter represent approximately one-third of that used to evaluate the merits of the national models (see Chapter III).
The chapter concludes with summaries of the countries’ primary purposes for evaluating publicly-funded research, their basic units of assessment, their core methods, the key indicators that they use to assess their publicly-funded research, researchers, and institutions, and classifications of their research evaluation system and funding mechanisms.

Of the 272 nations, dependent areas, and other entities in the world, this sample represents more than two-thirds of the world's top purchasing power parities, as well as a

- AU = Australia; BE = Belgium; CZ = Czech Republic; FI = Finland; FR = France; DE = Germany; HK = Hong Kong; HU = Hungary; IE = Ireland; JP = Japan; NL = Netherlands; NZ = New Zealand; PL = Poland; SE = Sweden; UK = United Kingdom; and US = United States.
large majority of the world’s ‘research superpowers’ in terms of their scientific productivity and government monies dedicated to research (DEST, 2003; European Commission, 2003; Group of Eight, 2002; Ministry of Research, Science, and Technology, 1999, 2001, 2005; Statistics New Zealand, 2006; Webometrics Ranking of World Countries, 2006) as evident in the international research landscape.\footnote{Almost no information was available for any of the African or South American nations.}

The International Research Landscape

A basic summary of the international research ‘landscape’ is presented in Table 1.\footnote{These data were obtained from DEST (2003), European Commission (2003), Group of Eight (2002), Ministry of Research, Science, and Technology (1999, 2001, 2005), and Statistics New Zealand (2006).} In this table, research intensity is a country’s gross domestic expenditure on R&D (GERD) as a percent of GDP for 2001 (in 2003 for New Zealand) ($M = 1.93\%$, $SD = 1.05\%$). Research budget is a country’s government budget allocated to R&D as a percent of GDP for 2003 (in 2001 for Australia) ($M = 0.66\%$, $SD = 0.25\%$). Budget growth rate is government budget allocated to R&D average annual growth rate in percent for the period 1997-2003 (from 1999-2003 for New Zealand) ($M = +3.56\%$, $SD = 3.60\%$).

Researchers is the number of full-time equivalent (FTE) in thousands for 2001 (2004 for New Zealand) ($M = 183.33$, $SD = 333.95$). Researcher growth rate is the average annual growth rate of FTE researchers in percent for the period 1996-2001 (from 1999-2004 for New Zealand and from 1981-2000 for Australia) ($M = +6.78\%$, $SD = 6.65\%$). Researchers/labor force is the number of FTE researchers per 1,000 labor force for 2001 ($M = 6.43$, $SD = 3.02$).

Research expenditure is the amount of money a typical researcher costs in a given country in $1,000$ United States dollars (USD) per FTE researcher in 2001 (in 1999 for New Zealand) ($M = $156.93 USD, $SD = $82.33 USD). This figure includes salaries,
equipment, materials, and services. Labor costs represent about 60% of the total.

*Publications* is the number of scientific publications in thousands per million population in 2002 ($M = 744.29$, $SD = 425.72$). *Publication growth rate* is the percentage growth rate for scientific publications for the period 1995-2002 (from 1991-1999 for New Zealand) ($M = +2.62\%$, $SD = 1.69\%$). *World publications* is national share of publications for the period 1998-2002 ($M = 7.15\%$, $SD = 9.65\%$). In total, for those nations for which data were available, these 16 nations produced nearly 80% of the world's total scientific publications (indexed by ISI) during the period of 1998-2002.

*Patents* is inventions patented at all three major patent offices—European Patent Office (EPO), United States Patent and Trademark Office (USPTO), and Japan Patent Office (JPO)—per million population in 1998 ($M = 43.29$, $SD = 37.29$).

These data are intended to set the tone and context for later comparisons of the national systems, and as shown in Table 1, the resources available in these countries for conducting research vary widely.

Additionally, the normative characteristics of each country's research evaluation model in terms of the primary reasons and motives for evaluating research, the model's basic units of assessment, its core methods, its key indicators and criteria, classification according to the model's systemization and consistency (i.e., general evaluation strategy), and its general funding model archetype are presented and described at the end of this chapter.
<table>
<thead>
<tr>
<th></th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research intensity</td>
<td>1.40%</td>
<td>2.10%</td>
<td>1.30%</td>
<td>2.50%</td>
<td>3.40%</td>
<td>2.20%</td>
<td>0.50%</td>
<td>0.90%</td>
<td>1.10%</td>
<td>3.00%</td>
<td>1.90%</td>
<td>1.10%</td>
<td>0.60%</td>
<td>4.20%</td>
<td>1.80%</td>
<td>2.80%</td>
</tr>
<tr>
<td>Research budget</td>
<td>0.30%</td>
<td>0.60%</td>
<td>-</td>
<td>0.80%</td>
<td>0.90%</td>
<td>1.00%</td>
<td>0.60%</td>
<td>-</td>
<td>0.30%</td>
<td>0.70%</td>
<td>0.70%</td>
<td>0.50%</td>
<td>0.30%</td>
<td>0.90%</td>
<td>0.60%</td>
<td>1.00%</td>
</tr>
<tr>
<td>Budget growth rate</td>
<td>-0.2%</td>
<td>+3.4%</td>
<td>-</td>
<td>+0.5%</td>
<td>+1.0%</td>
<td>+3.2%</td>
<td>+8.8%</td>
<td>-</td>
<td>+12.3%</td>
<td>+4.7%</td>
<td>+1.8%</td>
<td>-0.1%</td>
<td>+2.2%</td>
<td>+5.7%</td>
<td>+1.1%</td>
<td>+5.5%</td>
</tr>
<tr>
<td>Researchers</td>
<td>-40</td>
<td>14</td>
<td>259</td>
<td>36</td>
<td>172</td>
<td>11</td>
<td>14</td>
<td>8</td>
<td>675</td>
<td>42</td>
<td>15</td>
<td>56</td>
<td>45</td>
<td>157</td>
<td>1,216</td>
<td></td>
</tr>
<tr>
<td>Researcher growth rate</td>
<td>+5.4%</td>
<td>+7.2%</td>
<td>+2.9%</td>
<td>+2.4%</td>
<td>+8.6%</td>
<td>+2.6%</td>
<td>+13.3%</td>
<td>+7.1%</td>
<td>+7.3%</td>
<td>+1.8%</td>
<td>+5.1%</td>
<td>+29%</td>
<td>+1.6%</td>
<td>+5.6%</td>
<td>+4.3%</td>
<td>+4.2%</td>
</tr>
<tr>
<td>Researchers/labor force</td>
<td>6.9</td>
<td>2.9</td>
<td>6.5</td>
<td>13.7</td>
<td>6.5</td>
<td>2.7</td>
<td>3.6</td>
<td>4.9</td>
<td>9.1</td>
<td>5.2</td>
<td>7.8</td>
<td>3.2</td>
<td>10.1</td>
<td>5.4</td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td>Researcher expenditure</td>
<td>$182</td>
<td>$65</td>
<td>$237</td>
<td>$148</td>
<td>$214</td>
<td>$100</td>
<td>$47</td>
<td>$165</td>
<td>$252</td>
<td>$221</td>
<td>$38</td>
<td>$27</td>
<td>$270</td>
<td>$172</td>
<td>$216</td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td>1</td>
<td>929</td>
<td>415</td>
<td>731</td>
<td>1,309</td>
<td>712</td>
<td>-</td>
<td>374</td>
<td>647</td>
<td>550</td>
<td>1,093</td>
<td>-</td>
<td>266</td>
<td>1,598</td>
<td>1,021</td>
<td>774</td>
</tr>
<tr>
<td>Publication growth rate</td>
<td>+1.6%</td>
<td>+3.8%</td>
<td>+1.5%</td>
<td>+3.5%</td>
<td>+1.6%</td>
<td>-</td>
<td>+2.4%</td>
<td>+6.1%</td>
<td>+3.0%</td>
<td>+1.6%</td>
<td>+5.4%</td>
<td>+3.7%</td>
<td>+0.9%</td>
<td>+0.9%</td>
<td>+0.7%</td>
<td></td>
</tr>
<tr>
<td>World publications</td>
<td>2.80%</td>
<td>1.30%</td>
<td>8.70%</td>
<td>0.90%</td>
<td>6.30%</td>
<td>-</td>
<td>-</td>
<td>9.30%</td>
<td>2.50%</td>
<td>-</td>
<td>1.20%</td>
<td>2.00%</td>
<td>9.30%</td>
<td>34.40%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Patents</td>
<td>14.4</td>
<td>37.2</td>
<td>1.0</td>
<td>69.9</td>
<td>74.9</td>
<td>35.0</td>
<td>113.6</td>
<td>2.3</td>
<td>11.6</td>
<td>81.0</td>
<td>49.8</td>
<td>9.8</td>
<td>0.3</td>
<td>107.4</td>
<td>31.2</td>
<td>53.3</td>
</tr>
</tbody>
</table>
Historical Overviews, Research Structures and Contexts, and Evaluation Models

Although it is not the intent of this chapter to provide a complete account of each country’s research structures, policies, and evaluation procedures, important details regarding both the funding and evaluation of government-sponsored research are presented in order to provide a context within each country (Yin, 1994). Moreover, much of this information serves as the basis for rating the merits of these countries’ research evaluation models in subsequent chapters.

The description of each country includes three main parts: (i) a historical overview; (ii) the research structure and context; and (iii) the evaluation model. The historical overview provides a brief summary of major events that have occurred in the country and which are intended to set the tone for the other two parts (e.g., the unification of Western and Eastern Germany in 1990, the division of Czechoslovakia into the Czech Republic and Slovakia in 1993).

The research structure and context section provides a detailed description of past and present policies for funding research. The evaluation model section describes the principles and procedures used for evaluating publicly-funded research.

Both the research context and evaluation model sections are described in as much detail as was feasible, and were researched using government documents, scholarly literature, input from experts within the countries, and the European Commission’s ERAWATCH (a provider of information on national research policies, structures, programs, and organizations).
In the following, Australia’s research structure and context and evaluation model as well as a brief historical overview of the country is presented.

Historical Overview

Aboriginal settlers arrived on the continent, now known as Australia, from Southeast Asia about 40,000 years before the first Europeans began exploration in the 17th century. No formal territorial claims were made until 1770, when Captain James Cook took possession of the continent in the name of Great Britain. Six colonies were created in the late 18th and 19th centuries. Ultimately these colonies federated and became the Commonwealth of Australia in 1901. The new country took advantage of its natural resources to rapidly develop agricultural and manufacturing industries and to make a major contribution to the British effort in World Wars I and II. In recent decades, Australia has transformed itself into an internationally competitive, advanced market economy. It boasted one of the OECD’s fastest growing economies during the 1990s, a performance due in large part to economic reforms adopted in the 1980s. In 2006, Australia had a population of more than 20.2 million, most of which is concentrated along the eastern and southeastern coasts (Central Intelligence Agency, 2006).

Research Structure and Context

The organization, management, and funding of research in Australia has changed considerably since the 1980s, when a binary system, which differentiated between colleges of advanced education and universities, and funded research only for the latter, was abolished (Turpin, 2000). Since the introduction of the Unified National System
(UNS) in 1988, the number of universities eligible for public research funding in Australia has nearly trebled (Geuna & Martin, 2003), resulting in an increasingly competitive research environment, “both for individual researchers and institutions” (Turpin, 2000, p. 37).

Research funding in the 1990s was characterized by a dual-funding system comprised of performance-based block grants to institutions and direct research grants given to researchers or research centers. Block grants for research, research training and infrastructure were provided by the Department of Education, Training and Youth Affairs (DETYA) as part of their annual operating grant (Turpin, 2000). These funds were allocated on the basis of each institution’s research performance within the 36 universities of the national system, and were the primary source of public funding for research throughout the 1990s (Kemp, 1999; Turpin, 2000). The way these funds were converted into research practice “…was largely left to the universities themselves” (Turpin, 2000, p. 38).

Evaluation Model

In 1990 the Relative Funding Model (RFM) was introduced, where research support was measured by the Research Quantum (RQ), initially based on successful competition for Commonwealth Competitive Grants (CCG). Although, “when it was recognized that this did not fully represent research performance, the criteria were broadened to include other sources of funding” (Geuna & Martin, 2003, p. 293) additional measures including publications, for example, were incorporated into the formula.

In 1993, the Minister for Education announced that effective in 1995, the RQ would be allocated on the basis of a new Composite Index (CI). From 1998, DETYA
and subsequently, the Department of Education, Science and Training (DEST), would be responsible for gathering data for the CI, for calculating research funding allocations, and for advising the Minister on the CI. At that time the CI consisted of (Geuna & Martin, 2003, p. 294):

1. Research inputs (funding)
   a. Each university's funding from Commonwealth competitive grants
   b. Other public-sector research funding
   c. Industry and other research funding
2. Research outputs
   a. Scholarly publications by staff and students
   b. Higher degrees (Masters and PhDs) completed

The weighting of the CI components have varied from year-to-year. For example, if a university's share (of the total research activities for all universities) of national funding were 4.5%, its share of publications were 3.6%, and its share of higher degree completions were 5.3%, averaged over two years (Geuna & Martin, 2003), then its CI (assuming weights of 80% for funding, 10% for publication, and 10% higher degree completions) would have been 4.5% of the total RQ allocation available as shown in Equation 1.

\[
CI = \sum [(0.045 \times .80) + (0.036 \times .10) + (0.053 \times .10)]
\] (1)

In 1994, the Australian Vice-Chancellor's Committee proposed that the RQ incorporate qualitative elements, as the RQ had long been criticized for being overly mechanistic, and research performance was being evaluated solely on the basis of volume, not quality (Bourke, 1997; Geuna & Martin, 2003). In 2004, the Minister announced that the Australian Government would develop the basis for an improved
assessment of the quality and impact of publicly funded research and an effective process to achieve this (DEST, 2005), known as the Research Quality Framework (RQF). The RQF was to be transparent to government and taxpayers so that they are better informed about the results of the public investment in research; ensure that all publicly funded research agencies and research providers are encouraged to focus on the quality and relevance of their research; and avoid a high cost of implementation and imposing a high administration burden on research providers (DEST, 2005).

The RQF consultation process, completed in 2006, included workshops, Expert Advisory Group (EAG) meetings, consultation forums, a national stakeholder forum, and an Issues Paper (Commonwealth of Australia, 2005a) which sought to “set out a series of propositions which seek to focus stakeholders’ thinking on the main issues that need to be addressed for the development of an RQF: its design and subsequently its implementation” (p. 3), among others. More than one-hundred fifty submissions in response to the Issues Paper (e.g., Group of Eight, 2005) were received and used by the EAG for developing the RQF.

Coinciding with the RQF development activities were a number of studies intended to support that development, including, but not limited to, the Allen Consulting Group’s study Measuring the Impact of Publicly Funded Research which was to: (i) “produce a classification, or typology, of the benefits of publicly funded research taking into account international and Australian efforts to date to build such a typology” and (ii) “propose ways of systematically and cost-effectively measuring each of the benefits in this typology, taking into account performance monitoring and funding arrangements that impact on these measures” (Commonwealth of Australia, 2005b, p. 1), and as a result a number of working group papers were produced, including for example, International Benchmarking (Commonwealth of Australia, 2005c), Mechanisms of Assessment—Cross-
Disciplinary Research (Commonwealth of Australia, 2005d), and A Review of Current Australian and International Practice in Measuring the Quality and Impact of Publicly Funded Research in the Humanities, Arts and Social Sciences (Donovan, 2005).

The Research Quality Framework: Assessing the Quality and Impact of Research in Australia—The Preferred Model (Commonwealth of Australia, 2005e) was officially released by the Minister in 2005, which presented the features of research assessment which would underpin the forthcoming RQF. Although touted as the preferred model, “it is not the final model,” and has been surrounded by controversy as the Minister’s introduction to The Preferred Model announced that the ARC and National Health and Medical Research Council (NHMRC) will also be subject to the RQF which was “something never discussed with the academic constituency and subsequently highly resisted” (C. Donovan, personal communication, February 13, 2006).

Once implemented, the RQF “will provide the Australian Government with the basis for redistributing research funding to ensure that areas of the highest quality of research are rewarded” (Commonwealth of Australia, 2005e, p. 3), which will involve all of the Institutional Grants Scheme (IGS) and at least 50% of the Research Training Scheme (RTS). The definition to be adopted for research for the RQF proposed by the EAG is:

...a broad notion of research activity consistent with the definition of research and experimental development (R&D) as comprising creative work undertaken on a systematic basis in order to increase the stock of knowledge...[classified] into four types of activity: pure basic research; strategic basic research; applied research including new ways of achieving specific and predetermined objectives such as clinical practice; and experimental development including creative works and performance insofar as they directly relate to original basic and applied research (Commonwealth of Australia, 2005e, p. 7).
The EAG recommended the following (see Table 2) ‘possible’ scale for rating research quality (Commonwealth of Australia, 2005e),\textsuperscript{266} although it is still being debated (Commonwealth of Australia, 2005f):

Table 2

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>The majority of research outputs were considered to be in at least the top 20% of research in its field internationally, with a significant percentage (&gt;25%) in the top 10%. There was evidence of high international peer esteem and significant impact on the international academic community.</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>The research is unclassified.</td>
</tr>
</tbody>
</table>

In a like manner, the EAG recommended the following (see Table 3) ‘possible’ research impact rating scale (Commonwealth of Australia, 2005e):

Table 3

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Fundamentally altered policy or practice in a particular field, or produced a major, identifiable social, economic or environmental change, locally or internationally.</td>
</tr>
<tr>
<td>Moderate</td>
<td>Significantly altered policy or practice in a limited field, or produced an identifiable social, economic or environmental change.</td>
</tr>
<tr>
<td>Limited</td>
<td>Little or no identifiable change in policy or practice, and little or no identifiable social, economic or environmental change.</td>
</tr>
</tbody>
</table>

\textsuperscript{266} Mid-level definitions (2-4 on the quality rating scale) of quality have not yet been defined for the RQF (Commonwealth of Australia, 2005e).
It has been proposed that the ratings on the quality and impact scales be aggregated using a Research Quality and Impact Matrix (Commonwealth of Australia, 2005e), where research groups that obtain high quality ratings (e.g., > 2) and a high impact rating are given an additional rating point. There were discipline-based workshops throughout 2006, and an RQF implementation group is still to be formed, which will likely make the final decisions about specifics such as the quality and impact scales and whether or not the scales will be aggregated. Furthermore, “no one as yet has any real notion of how to measure impact,” and “the Group of Eight want to focus on quality,” whereas “the technical universities want to focus on impact.” (C. Donovan, personal communication, February 13, 2006).^267

Belgium

In the following, Belgium’s research structure and context and evaluation model as well as a brief historical overview of the country is presented.

Historical Overview

Belgium became independent from the Netherlands in 1830, and it was occupied by Germany during World Wars I and II. The country prospered in the past half century as a modern, technologically advanced European state and member of the North Atlantic Treaty Organization (NATO) and the European Union (EU). Tensions between the Dutch-speaking Flemings of the north and the French-speaking Walloons of the south have led in recent years to constitutional amendments granting these regions formal recognition and autonomy. Belgium had a population of 10.3 million in 2006 (Central Intelligence Agency, 2006).

^267 See also Donovan (2008).
Research Structure and Context

Belgium is unique amongst all EU Member States, namely that it is the only country where research policies are fully decentralized across several governments, enjoying complete autonomy of decision and power in these matters. The law states that the primary jurisdiction for research policy is within the three Regions and three Communities, while the federal state retains some competences as an exception to this rule. Governmental responsibilities are arranged as follows (European Commission, 2006a):

1. The Regions (Flanders, Wallonia, and Brussels-Capital) have authority on research policy for economic development purposes, thus encompassing technological development and applied research.

2. The Communities (French-, Flemish-, and German-speaking) are responsible for education and fundamental research at universities and Higher Education Establishments.

3. The Federal State retains responsibility for research areas requiring homogenous execution at the country level, and research in execution of international agreements (e.g., space research).

This institutional context has a profound influence on the governance of research policy. There are formally seven independent Belgian authorities (three Communities, three Regions, and the federal State) carrying out their own policy in the wider field of science, research, technology, and innovation. In practice, there are only six active entities, since the Region of Flanders and the Flemish Community's governments have merged into one entity. Due to its small size, the German-speaking Community does not carry out any policy in the research area. All the other entities have their own policies. There is no hierarchy of powers between the federal government and other authorities.
A large part of public research in Belgium is carried out in universities, which are major actors in the research landscape. They depend on the Communities for their funding and management, and are given a lot of autonomy. They also access funding sources from federal and regional levels. The third mission of universities—service to society in addition to training and research—is present, but not formalized in a law.

Belgium also has a well-developed network of collective research centers, dating from the 1950s, which have been established in the major industrial sectors, with industry financing. They have been the main actors for technology diffusion to companies in the past century. Regional research centers are also funded by the Walloon government, notably in the context of Structural Funds programs (Objective 1 funds in particular). The federal government funds targeted programs and administers a range of scientific institutes.

The Regions have established their own schemes for private R&D funding, ranging from subsidies, loans, small and medium-sized enterprises (SME)-specific grants, seed, venture capital, and equity schemes. Fiscal incentives for R&D do exist, and have recently been completed by a range of partial exemptions of advance payment on wages for employers who employ researchers, applicable in the public and the private sectors.

As mentioned, Belgium has a decentralized decision-making and governance system for R&D funding (CORDIS, 2007a). Within it, the Federal level accounts for a restricted part or the R&D funding (less than one-third). The Federal level provides funding for research of national interest such as security as well as for international research programs such as space research. Fiscal incentives are also present at the federal level.

Funding for basic research is the responsibility of the Communities. This includes both institutional funding, competitive funding across universities, and support
to individual researchers. Funding for applied and technological research with the view of increasing the value-added in the economy is the responsibility of the Regions. This includes both direct funding to companies and funding of research centers and technology diffusion activities. The relative shares of the federal and federated entities in overall budgetary credits for R&D in Belgium are as follows: 45.75% comes from the Flemish Community (including Community and Regional competencies); 29.63% from the federal state; 13.37% from the French Community; 10.23% from the Walloon Region, and 1.02% from the Region of Brussels-Capital.

Total budgetary allocations from the public sector for R&D in Belgium amounted to €1.7 billion ($2.3 billion USD) in 2005. These funds are allocated to the following sectors (European Commission, 2006a):

1. 26% as institutional funding to third-level education institutions and 12% for the large Research Funds which are (partly) allocated on a competitive basis
2. 13% to scientific institutions
3. 18% in the form of research action programs, which are open to public research and/or private research performers, also including individual grants for researchers
4. 11% for industrial research (there is also private research funding in the previous line)
5. 13% for international research programs

In terms of objectives pursued by the research funded by public credits, the following split is calculated (European Commission, 2006a):

1. 25% for non-oriented research at universities
2. 19% for the general funds at universities
3. 9% for socio-economic objectives
Enterprises are the main funding source of R&D activities in Belgium (60%). The public sector funds 23.5% of total R&D activities, and foreign sources account for 13% of the total funding sources. The mechanisms for funding fundamental research, under the responsibility of the Communities in Belgium, respect the principle of researcher's initiative and consequently do not incorporate any priorities in terms of sectors or disciplines. The same holds for the ‘inter-university attraction poles’ program of the federal authority, and other federal programs, which provide funding for research projects at universities, often in collaboration across the linguistic communities.

The French Community funds R&D in universities mainly through its basic allocation to universities, part of which is devoted to research. The Directorate General for non-obligatory education and scientific research implements and administers the policy for the Community. Additional funding for researchers and research teams, based on competition is channeled through the National Fund for Scientific Research (FNRS)—€95.0 million ($125.2 million USD)—and its associated funds, the Special Research Fund (SRF) and the Concerted Research Actions (ARC) program—€11.0 million ($14.5 million USD)—in 2005.

In Flanders, university research funding falls under the merged Flemish Government. For education and research in universities, the most important funding mechanisms are: the operational subsidies for the universities—€592 million ($779 million USD), the Special Research Fund (BOF)—€233 million ($306 million USD)—in 2005) and the Fonds Wetenschappelijk Onderzoek (FWO). FWO-Flanders (€115 million [$151 million USD] in 2005) is governed by the same Ministry and implements policy with regard to basic research at the universities. It is a funding channel which sets selection criteria and evaluation and decision-making procedures and distributes the funds amongst universities and research institutes (European Commission, 2006a).
The main aim of this funding for basic university research in both Communities is to finance fundamental research of high quality and to guarantee an excellent level of education for researchers. The policy is not thematically organized. Instead, the Communities leave the thematic choices to the researchers and focus on the quality of scientific research to support. There are three further principles to which the Communities subscribe, namely: promoting inter-university cooperation, promoting international mobility of researchers, and including research in the European Research Area (ERA).

A number of funds and public funding appropriations from Belgian authorities are dedicated to thematic priorities. At the Federal level there are thematic programs in areas which fall within the competences of the federal level, including for example space research (the most important in budgetary terms), and other federal research programs in areas such as information society, national cohesion, and normalization.

Wallonia has mobilizing programs which are a short-term research programs that are open either to universities and research centers, or to companies, and sometimes to combinations of private and public actors. These programs have, over the last ten years, notably covered biotechnology and nanotechnology. The areas for these mobilization programs are chosen amongst the ‘40 key technologies’ in which Wallonia has scientific and industrial expertise (European Commission, 2006a).

Flanders targets its R&D funding through the establishment (beginning in the 1990s) of major independent research centers, heavily supported by the regional government. Interuniversity MicroElectronics Center (IMEC) is considered the worldwide center of excellence in micro-electronics research. It is a unique and independent European research center in the field of nanoelectronics, nanotechnology, design methods and technologies for information communications technology (ICT)
systems. It has now become a world player with more than 1,400 researchers. Two other major centers are specialized in biology and environment and energy.

Collective research centers, dating from the 1950s and financed partly by industry, are present in all Belgian regions. With the federalization of the country, the public funding part of such centers has been placed under the responsibility of the Regions, which means that these public funds are indeed earmarked for a certain number of sectors. These sectors are defined according to the old industrial specialization structure (e.g., metalworking, textile, glass, ceramics, road construction, materials), but aim at developing and diffusing new technology applications into all industrial sectors. The Regions have also set up specialized ‘excellence centers’ in a number of sectors or technologies as well as other research centers.

Institutional funding in Belgium is mainly formed of funding to universities (basic allocations), institutional grants to research centers, and the functioning budgets of implementing agencies such as Institute for the Promotion of Innovation by Science and Technology (IWT), Directorate General for Technologies, Research and Energy (DGTRE), or Institute for the Encouragement of Scientific Research and Innovation (IRSIB) and the various science policy councils. There is no aggregated calculation of the specific shares of public R&D funding from all State entities devoted to this type of institutional funding.

Collaborative research with shared public and private funding of projects is common practice, but there is no aggregate source presenting the amounts of money involved across all Belgian authorities’ promotion schemes.
Evaluation Model

Evaluation practices vary widely across Belgian government entities and according to types of R&D instruments. However, the major research centers established in Flanders are regularly evaluated as part of their multi-annual agreement with the Flemish government. Such evaluations are carried out by independent experts, often from outside the country. Evaluations of various strands of activities of the main implementing agency, IWT are in the process. Several R&D programs evaluations have been organized by IWT, notably the SME aids scheme and the Flemish Cooperative Innovation Networks (VIS) scheme, and studies of the additionallity of R&D subsidies have been carried out by Flemish university departments or consulting companies. In general, the above-mentioned evaluations led to changes or redesign of activities in the institutes and programs concerned. IWT has established a monitoring and analysis (M&A) department, following the previous IWT Observatory, in order to prepare analyses of the Flemish innovation system and various aspects of it. Such studies, which are made publicly available, purportedly nurture policy-thinking and influence the design of new measures (European Commission, 2006a).

The federal authority orders external evaluations of some of its programs: the inter-university attraction poles program has been subject to an in-depth external evaluation, whose results have been incorporated in further calls under this program. An independent analysis of the R&D tax incentives system has also been carried out at the instigation of national authorities.

Evaluation of R&D performers or R&D programs is not widespread in Wallonia, though the strategic program Prométhée initiated a trend in this direction. During this exercise and as a follow-up of it, assessments of the R&D subsidies to companies and of the S&T intermediary system have been conducted, which spurred important debates in
policy circles and within the research community. The results of the latter study have been incorporated in the current policy framework—the ‘Marshall Plan’—featuring R&D policy for Wallonia.

The Flanders region of Belgium is the only region that has made a concerted attempt to introduce bibliometric methods into evaluating research performance. Six universities are involved, ranging from small to medium/large in size. FWO-Vlaanderen, the Fund for Scientific Research in Flanders, monitors a large portfolio of basic research grants and projects to individual researchers (including doctoral students and post-doctoral grants) and academic promoters at Flemish universities. The selection and monitoring mechanism is conducted by scientific commissions that base their decisions on a peer-review system, consistently involving foreign experts in evaluating the proposals submitted to the agency (European Commission, 2006a).

Besides the public R&D funding via FWO-Vlaanderen, which is distributed on a project-per-project basis or on an individual basis, the Flemish government created a mechanism that allows for supporting more large-scale basic research at universities. Except for setting certain quality guidelines and performance expectations, the government does not intervene at all in the internal selection and monitoring process for the grants. BOF had a total budget of €90 million ($119 million USD) to distribute across the six Flemish universities for fiscal year 2002. The weights were at first based purely on student numbers, according to a weighted scale (van Tunzelmann & Mbula, 2003).

The dissatisfaction with a numbers-led weighting is now being corrected by introducing explicit bibliometric indicators. This has led to the creation of a dedicated research and policy support staff, called Steunpunt O&O Statistieken (SOO), to support a major inter-university funding allocation decisions. Bibliometric data have for the first
time been used to allocate €93 million ($123 million USD) of public research money between these six Flemish universities for the fiscal year 2003, based on Web-of-Science (WoS) Science Citation Index (SCI) data provided to SOO via a license agreement with Thomson-ISI. Although the use of WoS data for evaluative and distributive purposes is not without controversy, they were considered the “best available, recurrently accessible, transparent and controllable” (von Tunzelmann & Mbula, 2003, p. 14) data for such a purpose. In addition patent data and innovation data are collected. Older criteria such as student numbers retain a 50% weighting in the overall allocation formula.

Moreover, as one of the architects of this system indicated in von Tunzelmann and Mbula’s (2003) *Changes in Research Assessment Practices*:

...he is partially critical of what it is currently trying to achieve. Many assumptions had to be made in preparing the data for comparison. Many of these are well recognized in bibliometrics work, such as problems of co-authorship across institutions, lead authorship, fractional authorship, self-citation, etc. Problems of misspellings and different listings of individuals and affiliations all had to be cleaned. The magnitude of this task was what led our correspondent to note that he could not see the feasibility of doing anything similar for a large region/country such as the UK.

Even then there are limitations on what is being achieved. Because of the particular limitations of SSCI and A&HCI, it is not being used for allocations in the social sciences or arts and humanities. Moreover the issue of research impact has not yet been fully addressed, as distinct from numbers of citations. Finally, our correspondent would wish the scheme to be extended to intra-university funding. He is however deeply concerned about over-use of any such tools, as likely to divert activity away from more academically useful research into tactics for cultivating citations (p. 15).
Czech Republic

In the following, the Czech Republic’s research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

Following World War I, the Czechs and Slovaks of the former Austro-Hungarian Empire merged to form Czechoslovakia. During the interwar years, the new country’s leaders were frequently preoccupied with meeting the demands of other ethnic minorities within the republic, most notably the Sudeten Germans and the Ruthenians (Ukrainians). After World War II, a truncated Czechoslovakia fell within the Soviet sphere of influence. In 1968, an invasion by Warsaw Pact troops ended the efforts of the country’s leaders to liberalize Communist party rule and create ‘socialism with a human face’ (Central Intelligence Agency, 2006).

Anti-Soviet demonstrations the following year ushered in a period of harsh repression. With the collapse of Soviet authority in 1989, Czechoslovakia regained its freedom through a peaceful Velvet Revolution. In 1993, the country underwent a ‘velvet divorce’ into its two national components, the Czech Republic and Slovakia. The Czech Republic joined NATO in 1999 and the EU in 2004. The Czech Republic had a population of 10.2 million in 2006 (Central Intelligence Agency, 2006).

Research Structure and Context

The present-day Academy of Sciences of the Czech Republic in its work continues the research traditions and mission not only of the former Czechoslovak
Academy of Sciences but also of its predecessors (Academy of Science of the Czech Republic, 2005). The oldest long-lasting learned society was the Royal Czech Society of Sciences (1784-1952) which encompassed both the humanities and the natural sciences. Among its founders were philologist Josef Dobrovský, historian Gelasius Dobner and mathematician and the founder of Prague University Observatory, Joseph Stepling; later it was headed by historian František Palacký.

By the end of the 19th century, language-differentiated scientific institutions arose in this country: the Czech Academy of Science and the Arts (1890-1952) and the Association for the Fostering of German Science, Arts and Literature in Bohemia (1891-1945) were established nearly simultaneously. The Czech Academy of Science and the Arts was founded owing to the significant financial support from the Czech architect and builder, Josef Hlávka who became its first President. The aim of this institution was to promote the development of Czech science and literature and to support Czech arts. The most important work of this Academy was its publication activities. Scholarships and financial support were also provided and smaller research units arose upon its initiative as well.

After the foundation of the independent Czechoslovak Republic in 1918 other scientific institutions were established, such as the Masaryk Academy of Labour and autonomous state institutes, such as the Slavonic, Oriental and Archaeological Institutes. Robust international relationships of Czech research institutions culminated in their affiliation with the International Union of Academies and the International Research Council.

After the totalitarian regime came to power in Czechoslovakia in 1948, all hitherto scientific non-university institutions and learned societies were dissolved and the Czechoslovak Academy of Sciences was founded (1953-1992), comprising both a
complex of research institutes and a learned society. Despite having been subjected to heavy ideological pressure until the fall of this regime in 1989, Czech science was nevertheless able to maintain its creative energy in a number of instances and to find its way to the world scientific community (although there were disparities with the various fields of sciences at different periods of the regime). This fact was made evident, among others, by the awarding of the Noble Prize to Jaroslav Heyrovský in 1959 and by the worldwide recognition attained by Otto Wichterle for his discovery of contact lenses (Academy of Science of the Czech Republic, 2005).

The Academy of Sciences of the Czech Republic (AS CR) was established as the Czech successor of the former Czechoslovak Academy of Sciences. It is set up as a complex of 51 research institutes and two service units including the Academy Head Office. The Academy employs about 7,000 employees more than a half of whom are researchers with university degrees.

The primary mission of AS CR and its institutes is to conduct basic research in a broad spectrum of the natural, technical, and social sciences, as well as the humanities. This research, whether highly specialized or interdisciplinary in nature, aims to advance developments in scientific knowledge at the international level, while also taking into account the specific needs of both Czech society and national culture. Scientists of the Academy institutes also participate in education, particularly through doctoral study programs for young researchers and by teaching at universities as well. The Academy also fosters collaborations between applied research and industry. The integration of Czech science into the international context is being promoted by means of numerous joint international research projects and through the exchange of scientists with counterpart institutions abroad.
The supreme self-governing body of the Academy of Sciences is the Academy Assembly; two-thirds of which is composed of representatives of all Academy institutes, the remaining third being representatives of universities, state administration, business circles, and other notable personalities. The executive body of the Academy is the Academy Council headed by the President of the Academy of Sciences. The Council for Sciences is primarily engaged in setting science policy of the Academy. Members of each of these Academy bodies are elected for a four-year-period. Academy Evaluation Committees, which correspond in their professional fields to respective science sections of the Academy, perform an independent assessment of the quality of research and research objectives of individual Academy institutes (Academy of Science of the Czech Republic, 2005).

The Academy of Sciences is financed primarily from the state budget. The pattern of research funding at the Academy conforms to current international standards. In addition to basic institutional financing of research objectives of Academy institutes, target-oriented financing is being more widely practiced to carry out research projects and grants selected on the basis of public competition. AS CR was the first institution in the Czech Republic to establish its own Grant Agency which financially supports research projects selected through a peer-review procedure involving reviewers from abroad. Individual Academy institutes obtain additional financial resources by participating in national as well as international research programs. The Academy has also been assigned financial responsibility for 71 specialized Czech scientific societies associated with the Council of Scientific Societies.

Total R&D expenditures (GERD) have been increasing in the Czech Republic since 1995 and in 2004 the total amount exceeded 35 million Czech Koruna (CZK; $263 million USD). The share of total R&D expenditures in GDP was 1.27% in 2004. Public
funds are stagnating and their share amounted to 0.53% of GDP in 2004.

The main provider of public R&D funding is the Ministry of Education, Youth and Sports which has a special position among other ministries in relation to publicly supported R&D. It coordinated the first National Research Program (NRP), finances research conducted at universities and provides institutional financing for research plans submitted by both public and private legal entities (as opposed to project financing on the basis of targeted research programs). Research and development in the Czech Republic is financed from public funds in two different ways (European Commission, 2006b):

1. Targeted financing through the support of research projects
   a. Grant projects proposed by individual researchers or legal entities
   b. Program projects fulfilling programs launched by providers
   c. Projects for state administration where the state is the only user of these results

2. Institutional financing through institutional financial support of research plans, specialized university research or international cooperation of the Czech Republic in the R&D field

The Czech Government has also introduced an indirect support of R&D through a new tax regulation which has been in force since 2005. This modification of the revenue act enables business entities to deduct expenses spent on R&D from their tax base. R&D related costs may be applied twice in the accounting—first as expenses as such and then separately as an amount to be deducted from the tax base before taxation.

Basic research is funded either through institutional financing provided to research entities or on the basis of grants provided to applicants on the basis of calls for research proposals. Basic research funding is allocated mainly to universities, Academy of Sciences, and other state research institutes. According to the OECD’s *Science, Technology*
and Industry Outlook (2004)—based on Czech Statistical Office data—the Czech Republic invested into the basic research more than 40% of the total R&D expenditures (GERD) in 2002. This is the largest share of GERD dedicated to basic research among the European Countries.

In 2005 the Ministry of Education, Youth and Sports launched a special program focused on the basic research—Centers of Basic Research. The main objective of this program is the support of cooperation between top level research establishments in the Czech Republic and the growth of their competitiveness in the European research area.

As a new input into the National Research and Development Policy, a set of seven Long-Term Principal Research Directions (LPRDs) was adopted by a Government Resolution in 2005—as an amendment to the National Research and Development Policy: Sustainable Development; Molecular Biology; Power Sources; Materials; Competitive Mechanical Engineering; Information Society; Security.

Targeted funds are distributed mainly through the National Research Program (NRP) II Thematic and Systemic Programs of the NRP II (2006-2011) as adopted by the Government in 2005. These include (European Commission, 2006b):

1. Sustainable prosperity
2. Quality life and environment
3. Information technologies for a knowledge-based society
4. Socio-economic development of the Czech Society
5. Human resources
6. International cooperation
7. Support to the preparation and implementation of a national policy including technical assistance.
Financing from public funds amounting to 7.3 billion CZK ($322 million USD) is planned for the NRPII for its whole duration, with 40% allocated to the Thematic Program on Sustainable Prosperity (TP1), 35% to the Quality Life & Environment (TP2), 12% to the IT for a Knowledge-based Society (TP3), less than 3% to the Development of the Czech Society (TP4), over 6% to Human Resources (PP1), 3% to International Cooperation (PP2), and 1% to support to the preparation and implementation of a national policy.

A share of 10%-20% of public R&D funding is to be allocated to the NRP and its priorities. More than 910 million CZK ($39 million USD) are to be allocated to the NRP in 2006. Calls for proposals on research projects are launched by the providers of financing (i.e., the Ministry of Education and the Ministry of Industry and Trade).

Targeted support is provided also through grants of the Czech Science Foundation (Grant Agency of the Czech Republic). Calls are not thematically focused. Institutional funding is provided mainly to the institutes of the Academy of Sciences, the Ministry of Education, Youth and Sports, and other ministries. These public R&D funds are mostly allocated dominantly to activities, from which no practical output is expected, such as basic/fundamental research and specific research at universities. Practical application of research results is not emphasized in these allocations—they are mainly just published (European Commission, 2006b).

One of the measures of the National Innovation Policy addresses this problem in that the National R&D Policy is to be updated so that the proportion of institutional R&D financing to financing of targeted/program-based R&D is shifted from current 60:40 to 40:60 by 2010. Also, the public support will be preferentially allocated to excellent research teams and emphasis will be placed on the evaluation of research results.
Institutional funding in the Czech Republic is higher than the EU average; that is, 60% of the R&D funding and targeted research funding is only 40% of the total amount. Public research funding could be up to 50% for industrial research and up to 25% for development. Public support may be increased by 10% of eligible costs if the applicant falls into the category of an SME, by 10% of eligible costs if research is related to a region having a low standard of living (complying to a definition of such a region), by 5% if it is directed at developing certain economic sectors as defined by the government, by 15% if it is R&D aimed at fulfilling objectives of the EU Framework Program (FP). Public support may amount in total to maximally 75% for industrial research and 50% for development (European Commission, 2006b).

**Evaluation Model**

In 2004, the Czech government approved a methodology for evaluation of R&D sponsored from public funds and R&D results. The methodology of R&D evaluation was modified in 2005 to improve several problematic points such as evaluating principles, evaluating effectiveness of R&D institutions, and evaluating R&D programs (European Commission, 2006b).

One of the four main objectives of the National Innovation Policy is to make the performance of state administration in research, development, and innovation more effective. It has been acknowledged that an effective support to R&D and innovation requires an efficient and coordinated state administration.

The Council for Research and Development has to produce regular annual analysis of the existing state of research and development in the Czech Republic and a comparison with the situation abroad. These materials have been submitted to the government which used the findings and recommendations stemming from analyses for
adopting acts and resolution dealing with R&D issues. Strategic documents and policies are elaborated differently. Impacts and effects of individual policies are still not evaluated. The evaluation of these strategic documents is formal and limited to administrative aspects of meeting government resolutions (European Commission, 2006b).

Before the establishment of the Grant Agency, a number of ministries, universities, and the Academy of Sciences began distributing some of their resources in the form of competitive grants. In all cases, the decision making process was much the same. Applications were solicited, then each application was subjected to one or more peer reviews and to quantitative assessment (Vrbova, 1997, p. 94).

As reported in the National Report on EU Governance Research: Czech Republic + Slovak Republic (2006):

A significant problem of Czech research in EU studies remains the unclear criteria for evaluation of the research outcomes. No formal evaluation system (analogous to impact factor system) exists in the Czech Republic. The first steps in the comparable evaluation of individual research institutions were made by the Institute of Economic Studies of the Faculty of Social Sciences of the Charles University in Prague—the evaluating scheme was, however, limited to research in economics (Šlosarčík, 2006, p. 9).

While peer review continues to be the predominant method of evaluating scientific research in the Czech Republic, it has been suggested that this system lacks objectivity and:

...leading positions in universities and research institutes are not always occupied by outstanding scientists or even highly qualified experts. This state of affairs results from the fact that, in the past, ideological and political criteria were frequently applied in the appointment process (Koutecky, 1997, p. 160).

Finally, as outlined in National Research and Development Policy of the Czech Republic for 2004-2008 (Ministry of Education, Youth and Sports of the Czech Republic, 2004) much
of the Czech Republic’s research evaluation is currently driven by the Lisbon Strategy, Barcelona objectives, and more recently, the EU’s Seventh Framework Programme (FP7).

There are no doubts that the Lisbon strategy belongs among the most important external factors influencing the formulation of the national R&D policy of the Czech Republic being the candidate country at the accession door to EU...It is aimed at creation of worldwide highly competitive economy within EU based on the society of knowledge and capable of steady growth at simultaneous creation of new jobs. The research supported from the state and private funds is the key factor at creation of new knowledge and plays a fundamental role at transition to an economy based on the society of knowledge...Then on the grounds of the Lisbon strategy and Barcelona objectives the European Commission presented an action plan to the European Council...for removal of the main weaknesses of research and development in the European countries. These weaknesses being understood as cause of hesitancy of investors to support research and development are common to all member and candidate countries of EU; the Czech Republic is no exception. Among the typical weak points are the shortages and low

---

268 The Lisbon Strategy is intended to deal with the low productivity and stagnation of economic growth in the EU, through the formulation of various policy initiatives to be taken by all EU member states. The broader objectives set out by the Lisbon Strategy are to be attained by 2010. It was adopted for a ten-year period in 2000 in Lisbon, Portugal by the European Council. It broadly aims to make Europe, by 2010, the most competitive and the most dynamic knowledge-based economy in the world.

269 Europe needs to spend more on R&D and technological innovation if its economy is to be as strong as, or stronger than, that of its main competitors. With this in mind, at the Barcelona European Council in 2002, EU leaders endorsed a target of 3% of gross domestic product (GDP) for overall R&D spending. The present average level is 1.9% of GDP, compared to 2.7% for the US and 3% for Japan.

270 Europe invests €120 billion ($160 billion USD) less in research than the US each year, and the gap is growing. Since the gap in public research spending is quite small, EU leaders set industry the target of coming up with two-thirds of R&D spending. To reach the Barcelona objectives, research expenditure in Europe will need to grow at an average annual rate of 8%, shared between a 6% growth rate in public expenditure and a 9% rate for private investment. Meeting the objectives is expected to increase GDP by 0.5% per annum after 2010 as well as creating 400,000 new jobs each year (CORDIS, 2003).

271 The EU’s FP7 states that “knowledge lies at the heart of the EU’s Lisbon Strategy to become the ‘most dynamic competitive knowledge-based economy in the world. The ‘knowledge triangle’—research, education and innovation—is a core factor in European efforts to meet the ambitious Lisbon goals. Numerous programmes, initiatives, and support measures are carried out at EU level in support of knowledge. The FP7 bundles all research-related EU initiatives together under a common roof playing a crucial role in reaching the goals of growth, competitiveness and employment; along with a new Competitiveness and Innovation Framework Programme (CIP), Education and Training programmes, and Structural and Cohesion Funds for regional convergence and competitiveness. It is also a key pillar for the ERA. The broad objectives of FP7 have been grouped into four categories: Cooperation, Ideas, People and Capacities. For each type of objective, there is a specific programme corresponding to the main areas of EU research policy. All specific programmes work together to promote and encourage the creation of European poles of (scientific) excellence” (CORDIS, 2007b).
flexibility of the professional career of the researchers driving out the top workers to abroad, as well as fragmentation and small distinguishability of the excellent research and problems faced by the top technological small and medium sized companies at obtaining the support for their research and innovations. And last, but not least weak point is the deficient knowledge of the research workers and managers in the area of intellectual property and putting of the R&D results into practice (Ministry of Education, Youth and Sports of the Czech Republic, 2004, p. 8).

In response, the Ministry of Education, Youth and Sports of the Czech Republic (2004) has outlined the following strategies for the evaluation of its publicly-funded research:

1. NR&DP will respond to the growing importance of the evaluation of research on all levels. At the same time it must struggle with the fact that the evaluation represents highly complicated and demanding activities necessary to be realized throughout the area of research and development on a high level and according to a single methodology.

2. The system of research evaluation in the Czech Republic will respect the global trends and employ new knowledge and best practice of research evaluation from the individual member countries of the EU or the OECD, respectively. The fundamental principals of evaluation will be the multicriterial approach, demonstrable professional competency, factuality, transparency, independence, and objectivity.

3. Particularly the higher quality of:
   a. The continuous evaluation of the process of implementation of NR&DP and its successfullness at desirable increase in the overall performance of Czech research and development
   b. The process of evaluation of the research results as an instrument of policy serving for more effective allocation of public funds to individual programs, projects, and institutions

4. The evaluation of NR&DP implementation, to include:
   a. Evaluation of implementation of the NR&DP all results and experiences will be utilized; but further intensification in the implementation, communication and methodological area is
expected. An increased attention will be given to the responsibility for fulfillment of the policy-related tasks (analysis of reasons of any failure, measures taken for the remedy), analysis of surviving problems and assessment of efficiency of individual policy instruments and their application under the given conditions. More independent professionals will become involved in the policy evaluation and its results will be presented to the broad public discussion. On the basis of the results of evaluation the Ministry of Education, Youth and Sport will consider any proposal of updating to the applicable national policy. Regarding the need to integrate to EU as a competitive member the indicators and evaluation instruments used in EU will be taken over. The results of evaluation of NR&DP implementation will become part of the analytical groundwork for preparation of the subsequent national policy.

5. The evaluation of programs and projects, to include:

a. Evaluation of the results of a research NR&DP puts emphasis on their originality, contribution to the overall knowledge, innovation and application benefits (with a view to specifics of the individual types of research) and to the whole range of social functions and effects of research for the economy, society, education, and knowledge.

b. Balanced attention given to the improvement of processes of both the ex ante evaluation of draft research programs and projects and the ex post evaluation that would demonstrate more deeply the actual quality and benefits of a solution and purposefulness of the public funds spent. The best evaluation of programs and projects requires a clear determination of research objectives and creation of a system of suitable and adequate criteria for the given type of research. During evaluation a due attention will be given to the feasibility criteria.

c. Evaluation of programs based upon evaluation (thematic, cross-sectional and partial programs) summarizing the expected, continuous, and final results (economic, social, and others) of the particular included projects. During solution and after termination of all projects included in the program the evaluation reports will be submitted to the government, including estimate of effectiveness of the financial means spent, by a particular provider or coordinator.

d. Evaluation of projects the whole evaluation scale will be used for
a reasonable differentiation in quality and successfulness of solution between individual projects and mineralization of a formal approach of evaluators. The importance of ex post evaluation will be increased to become not only the audit of allowable cost regardless of the results of the project solution. The results and knowledge of ex post evaluations will serve as feedback for future decision-taking on selection of projects and their investigators. The R&D information system will include relevant information on results of projects after the evaluation is ended.

c. Evaluation results (with emphasis given to the results of a long-term repeated evaluation) will become the decisive criterion for allocation of disposable financial means. Preferred will be those who are reaching better results in the long term.

d. With the aim to reach an overall higher quality of research evaluation the procedures used in the advanced countries will be analyzed and selected methodologies and indicators applied to Czech conditions. The Czech Republic will become more involved in the international system of evaluation and benchmarking will be applied to a greater extent. Also, case studies on the ‘fate’ of selected projects and research results in general, to increase the overall evaluation culture. At the same time the development of the professional base for the area of evaluation will be supported, as well as other relevant issues of the research policy.

6. The evaluation of research institutions and individuals, to include:

a. All research institutions spending institutional funds will be regularly and thoroughly evaluated by providers of these institutional funds, or by their establishers, with participation of foreign experts and representatives of the central bodies of the state administration. Also, to dedicate attention not only to the effective spending of the obtained public funds, but to the ability of these institutions to obtain and increase the value of the private funds on the basis of their results.

b. The evaluation of individual research workers is particularly an internal matter of development of the research institutions themselves. To be efficient, it must have a motivational character and evaluation must be connected with the career order and made on regular basis. Besides the quality of the research work itself, beyond dispute being the most important criterion, also
engagement in international research projects and teams will be evaluated. Besides publication activity (evaluated by prestige of particular periodicals and publishers), application outcomes (e.g., patents) from the private sphere will gain increased importance. The results of evaluation will be used by particular providers as the groundwork for determination of the overall amount of the institutional support.

It is not clear, however, whether these strategies have been (or will be) put into practice in the Czech Republic.

Finland

In the following, Finland's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

Finland was a province and then a grand duchy under Sweden from the 12th to the 19th centuries and an autonomous grand duchy of Russia after 1809. It won its complete independence in 1917. During World War II, it was able to successfully defend its freedom and resist invasions by the Soviet Union—albeit with some loss of territory. In the subsequent half century, the Finns made a remarkable transformation from a farm and forest economy to a diversified modern industrial economy. Finland’s per capita income is now on par with Western Europe. As a member of the EU, Finland was the only Nordic state to join the Euro system at its initiation in January 1999. In 2006, Finland had a population of 5.2 million (Central Intelligence Agency, 2006).
Research Structure and Context

In Finland, research councils give grants based upon international peer review, and governments give institutional funds. A new budgeting system was introduced in 1994, in which a distinction is made between funds for teaching and for research. The budget contains five elements: a basic grant; a performance-related grant for teaching; a research grant; a grant for other activities; and a capital grant. Until 1995, research funds were allocated on an incremental basis. Since then, amounts awarded have depended upon the volume of teaching and external research income. No other performance measures are used, although doctoral student numbers help determine the performance-based grant for teaching.

Most academic research in Finland is conducted in the Finnish universities, and is financed by core funding from the Ministry of Education; by peer-reviewed grants awarded by the four research councils under the Academy of Finland; and by research contracts from industry and government. The Ministry of Education uses a system known as Management by Results, introduced in 1994, in which a small proportion of every budget is based on an assessment of performance. This was followed by the adoption in 1998 of three-year agreements that specify the outcomes that each university is expected to achieve, and the levels of funding that each receives. These three-year agreements are updated annually.

At present, the agreed areas of expenditure comprise basic (90%), project (7%), and performance-related (only 3% at present, but expected to increase). Basic funding covers salaries and facilities. A formulaic model for basic funding has been used since 1997, in which a connection between teaching and research is made explicit. The teaching component is represented by target numbers of Master's degrees and the
research element by target numbers of doctoral degrees. Project funding is earmarked for programs that the government defines. Performance-related funding is awarded on the basis of quality and impact indicators, and has been used to establish centers of excellence, to increase international collaboration, to improve graduate placement, and to meet planning targets (European Commission, 2006c).

Currently, approximately 90%-92% of the budgetary allocation to universities is based on the number of masters and doctoral degrees produced at the universities. The doctoral degrees (32%) are regarded as indirectly reflecting the university's research performance. Additionally, 3%-5% of the basic budget is allocated according to quantitative and qualitative performance indicators, including the centers of excellence, funding from the Academy of Finland, international funding, international exchange of teachers and researchers, graduate employment, and estimation of the renovations and development trends at university (Kaukonen, 1997).

The rest of the basic budgetary funding, 3%-5%, is reserved for specific project funding in the nationally-defined priority areas (including both research and educational projects) which are agreed on in the performance negotiations with the Ministry. The meaning of this sharing is to give the Ministry of Education the possibility of launching projects and supporting activities which are considered to be of nation-wide importance, and to give the universities a possibility for "new openings" in research, teaching and development work. At the moment, some examples of the largest projects are the Graduate (doctoral) Schools and projects related to the advancement of the Information Society, both in research and teaching (Kaukonen, 1997).
Evaluation Model

The Academy of Finland has been evaluating research performance since the 1970s, focusing mainly on individual scientists, projects, and teams. Initially, there was little by way of a systematic, nationwide ex post evaluation of research. However, beginning in the 1980s, there have been evaluations of about twenty fields, including inorganic chemistry (1983), automation technology (1986), legal science (1994), and molecular biology and biotechnology (1997). These have focused on international outcomes and have been driven more by a desire to improve the quality of science than by the need to make funding decisions (European Commission, 2006c).

The Finland Higher Education Evaluation Council (FINHEEC), established in 1995, conducts evaluations of three main kinds: institutional; program/thematic; and accreditation. None of these is targeted specifically at research, however. Nor do institutional evaluations use a uniform model for all universities. On the contrary, the government recognizes differences between universities and emphasizes the developmental role of evaluation. As a result, most evaluations are broad assessments of basic preconditions for teaching and research and the capacity for change. These include statements of an institution’s mission, processes, institutional arrangements, resources, and performance. Less attention is paid to the latter two factors. Emphasis varies across universities; one might highlight its teaching, another, its regional role (European Commission, 2006c).

FINHEEC evaluations take place in three phases. First, each university carries out a self-evaluation and prepares a report, which is assessed by an external team that visits the university and then produces a final report. Academy evaluations, however, proceed differently. In the case of electronics research, for example, the Academy of
Finland commissioned an evaluation in 1985 from the Research Council of Natural Science and Engineering. A committee was set up and two international experts appointed. The scope was limited to certain pre-defined sub-areas, in which twenty-eight university groups were identified and evaluated with respect to: (European Commission, 2006c)

1. Mission, vision, and goals
2. Efficiency in using resources
3. Scientific competence and degree of innovativeness technological competence and cooperation with other researchers, industry, and users
4. Technological competence and cooperation with other researchers, industry, and users
5. The national/international importance of results for the scientific community
6. The relevance of a group’s research for industry

Evaluations were conducted in three phases. A questionnaire was distributed to groups and, having examined the results evaluators interviewed each group, summarizing their findings. Groups were given opportunity to comment. In a report entitled, Management by Result (2003), the Ministry of Education proposed a performance-based mechanism similar to that of the United Kingdom’s RAE, and suggested that 35% of funds should be allocated on the basis of research performance. It advocated that all university groups be evaluated by the Academy of Finland every three years, using peer review, with research units being graded on a five-point scale, which should be used to determine the funds they receive. The suggestion was criticized by almost all the universities, and the proposal was ‘frozen’ by the Ministry. The main objection was that the mechanism would give the Academy undue influence.
The Academy of Finland has carried out policy and program evaluations mainly on two levels: at the level of research programs and at the level of research fields. All research programs of the Academy of Finland are evaluated. Research programs are evaluated against the starting-points of the program, its objectives, and funding volume. The main focus is on the performance of the program as a whole as well as on the added value it has generated, but evaluations are also carried out at the level of individual thematic areas and projects. Research program evaluations typically focus on the following issues (Academy of Finland, 2003):

1. Scientific results of the research program
2. Impacts of the research program
3. Implementation of the research program
   a. Preparation and planning of the contents of the research program
   b. Funding decisions and coordination

As for the evaluations of research fields, the focus is typically on the following issues (European Commission, 2006c):

1. Scientific quality of research carried out by Finnish organizations
2. Scientific relevance of future research plans
3. The appropriateness of research methods
4. Sufficiency of resources
5. Capacity of research groups and organizations
6. Interaction with the international scientific community

Research Councils, the Academy's Board, and other funding bodies supporting research have a key part to play in utilizing the results of evaluations. Research Councils make use of the recommendations made on the strength of the evaluations. Based on the recommendations, they draw up an after-care plan for the program, including
proposals on the implementation of the recommendations in so far as they are considered justified.

Evaluation activities are extensively and systematically carried out in Finland. Evaluations have focused on research programs, research organizations, research funding organizations, as well as to some extent on research policy. During recent years, the importance of evaluating the impacts of research has increased. This trend is evident both in applied as well as in basic research. The underlying reason for this development is the increased steering of public sector organizations based on agreed performance indicators as part of their annual performance agreements. There are increasing pressures also for research organizations to generate evidence of impacts of their activities.

As far as evaluation processes are concerned, FINHEEC does not have a rigid, predetermined evaluation pattern that is applied to every project. In fact, the chosen methods can vary according to the target of evaluation and phrasing of evaluation questions. In the beginning of each project, the objectives and implementation practices of the evaluation are defined. The point of view and suitable evaluation methods for the project are also chosen at this stage. When deciding both on the process and methods, FINHEEC actively takes the special characteristics of the evaluation target into consideration in order to achieve a coherent evaluation scheme that corresponds to the evaluation needs in the best possible way. This includes analyzing the perspectives of the higher education units under review.

The following outline, however, is a basic pattern that most evaluations follow (FINHEEC, 2007):

1. The Council makes a decision on an evaluation and appoints a steering committee

2. The steering committee makes a proposition to the Council about
the composition of an external evaluation team and prepares both a review and project plan

3. The Council appoints the external evaluation team and approves the project plan

4. The higher education institution(s) under review compile(s) self-evaluation reports to the external evaluation team

5. The external evaluation team visits the higher education institutions involved and writes a review report

6. The review report is published

Another trend evident in Finnish evaluation practice is the shift from evaluation of individual research projects to impact evaluation at the program level. Tekes (the main public funding organization for research and development in Finland; Tekes, 2007) has been particularly active in initiating evaluations that focus on the added value of program-level activities and services (European Commission, 2006c).

Finnish evaluations of research can be considered as a continuous process that can be applied at different phases of a research cycle: before, during, and after. Evaluation can also cover all levels of R&D activity, from the macro-level of the national R&D system downward to the various micro-level components. Methods for different levels of evaluation, for different stages, and for different types of research activity have been developed and used with varying degrees of reliability.

Evaluation of research projects or individual researchers has been conducted beforehand for all research plans submitted for support to the Academy of Finland. On the project level, evaluations of on-going research are also common; for example, by monitoring groups of experts appointed for major projects by the Academy of Finland. Ex post evaluations take place on the level of individual researchers when competence and achievements are evaluated, for example in connection with academic appointments (Kaukonen, 1997).
The evaluations have relied heavily on the classical peer review. However, "where the national scientific community is relatively closed and small—as in Finland—the well known problems of objectivity associated with (domestic) peer review become magnified" (Kaukonen, 1997, p. 18). The solution adopted is to draw peers from the international community. Thus, almost all the evaluations have been conducted by an international evaluation group. The major exception has been the evaluation of the science of education where the choice of experts was limited to Finnish researchers only.

The disciplinary evaluations have emphasized the assessment of research excellence in terms of international prestige and contribution to the forefront of science. It has become obvious, though, that the reliance on international experts in evaluations has limited potential in the applied and social sciences and humanities. In these fields, where publications are mostly in the native language, there are not many competent foreign peers, and a variety of criteria of assessment are needed. As Kaukonen (1997) has noted regarding the use of external peers:

The use of international experts may not be unproblematic even in relatively uncontroversial research areas, since foreign peers are usually confronted with the current international state of affairs while lacking knowledge about the historical, structural and organisational context of research activities in Finland. An equally important issue is that the use of international experts is explicitly connected with the criteria of assessment in terms of "comparison to the international top level". The relative position of a field in this regard, if taken as an exclusive criterion of evaluation, is only applicable to a few non-controversial basic research areas (Kaukonen, 1997, p. 18).

The standard procedure in these evaluations is that an evaluation report is prepared collectively by invited foreign experts, based on summaries of research activities and publications of research groups, site-visits, and interviews with scientists. With some differences, the assignments of the evaluation groups have consisted of the evaluation of the sufficiency and appropriateness of research posts, equipment and other
resources, the quality of research, and the future plans for research development. Depending on the substance, the research conducted in Finland has also been compared with corresponding research conducted in other countries. All the disciplinary evaluations have also included evaluation of post-graduate education (training of researchers) in respective fields. The data gathered for the evaluation have consisted of scientific publications, statistics of higher education and research, general overviews of the research organization, its staff and activities, funding profiles and descriptions of future plans. The evaluation reports also involve detailed descriptions of individual departments. The Academy of Finland has covered the evaluation costs which, on average, have amounted to 300 thousand Finnish Markka (MK; $79 thousand USD) per evaluation.

The new evaluation practices started to influence the everyday life of university-based researchers more profoundly in the beginning of 1990s as the evaluation boom gained new momentum. This was related to new demands on university activities, research included, to be more accountable, efficient and to produce 'top results' according to international standards. These demands are now reinforced by real threats of cutting university funding, on the one hand, and by the incentives of getting extra resources for good performance, on the other.

**France**

In the following, France's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.
Historical Overview

Although ultimately a victor in World Wars I and II, France suffered extensive losses in its empire, wealth, manpower, and rank as a dominant nation-state. Nevertheless, France today is one of the most modern countries in the world and is a leader among European nations. Since 1958, it has constructed a presidential democracy resistant to the instabilities experienced in earlier parliamentary democracies. In recent years, its reconciliation and cooperation with Germany have proved central to the economic integration of Europe, including the introduction of a common exchange currency, the Euro, in 1999. At present, France is at the forefront of efforts to develop the EU's military capabilities to supplement progress toward an EU foreign policy. In 2006, France had a population of 60.8 million (Central Intelligence Agency, 2006).

Research Structure and Context

Following the implementation of the National Agency for Research (ANR; Agence Nationale de la Recherche) in 2005, a share of the government funding of public basic research is now allocated through an intermediary research agency. In 2005, the budget of the ANR reached €350 million ($461 million USD). The majority of funding (80% in 2005) was dispensed through calls for project proposals. The remaining 20% were distributed for the specific actions for which the State had committed itself (European Commission, 2006d).
Evaluation Model

During the 1970s and 1980s, under the General Directorate for Scientific and Technical Research (DGRST; Direction Générale de la Recherche Scientifique et Technique), research in France was seen as dynamic and productive and as having an excellent reputation. At this time, in 1984, the National Committee for Evaluation (CNE; Comité National d’Evaluation) was established to evaluate France’s universities, including their research, and in 1985 a bill was passed that required all research institutions to establish structures for evaluating their work. In 1989, a decree was made which resulted in the creation of National Committee for Research Evaluation (CNER; Comité National d’Evaluation de la Recherche). This situation continued during the early years of the Directorate for Research and Test Facilities (DRME; Direction des Recherchés et Moyens d’Essais), a recognized military institution much admired by civil researchers. From 1986 to 2003, structures and procedures for assessment at different levels (i.e., individual assessment, research unit assessment, and program and institution assessment) were created (Barret, 2004, June).

Today, the situation has changed and “French research is in bad shape” (Comité National d’Evaluation de la Recherche, 2006, p. 13). These difficulties came to head in 2004, when researchers began a protest movement and sought to inform the government and their fellow citizens on the negative consequences of dysfunctions in the French research system. This protest led to the creation of two collectives—Let’s Save the Research (Sauvons la Recherche) and the Committee for Initiative and Progress (CIP; Comité Initiatives et Progress)—and the holding of the Research Assizes. During these debates, the importance of, and need for, evaluation in orienting governmental policy for research and careers in research emerged as a major issue. As a result, CIP proposed the
merger of France's two evaluation bodies: National Evaluation Committee for Public Scientific, Cultural and Professional Establishments and the CNER to form the Agency for the Evaluation of Research and Higher Education (AERES; Agence d'Evaluation de la Recherche et de l'Enseignement Supérieur). The Ministers for Education and for Research, together with the directors of these two bodies, drew up a strategy for cooperation based on CIP's proposals.

CNER, in collaboration with CNE, then prepared a report entitled *Creation of a Single Evaluation Body for Higher Education and Research*, which was submitted to the Ministers for Education and for Research in 2005 (Comité National d'Evaluation de la Recherche, 2006). This report identified three major issues: the importance of preserving the acquis of each committee and of benefiting from the dynamic they had created; consideration for the specificities of universities and for their autonomy; the importance of developing an analytical model, while avoiding recourse to academic criteria alone.

Prior to 2001, France had “no evaluation mechanisms which were specifically designed to evaluate university research” (von Tunzelmann & Mbula, 2003, p. 16). In response to this lack of agreement on evaluation techniques, the CNE published the *Guide d'Evaluation des Universités*, in which the procedures of evaluation are detailed. This evaluation is conducted by the CNE and has two phases: one internal and one external.

The internal evaluation is organized by the institution. This evaluation is helped by the guidelines for evaluation and must involve all the institution's staff. It consists of the preparation of an internal evaluation report that will be the main guideline for the external evaluation. Second, the CNE organizes and coordinates an external evaluation based on a peer review. The experts include university professors, higher education administrative or technical senior executives, and key economic professionals, be they French or not. The final evaluation report is prepared by the CNE on the basis of the
three main sources of information: the self-evaluation report; the external peer-review report; and the visits organized by the CNE. This final report has to be confirmed by the president of the evaluated institution.

CNER's procedures are similar to those of CNE and consist of four stages (Comité National d'Évaluation de la Recherche, 2007):

1. Identification of the field, the specifications, and an explanation of the approach to operators
2. Selection of French or international experts, working-out contracts, submission of one or more special reports and a combined report communicated as a directive report
3. The plenary committee of CNER takes note of observations, recommendations, or reserves from the decision-makers of the evaluated bodies
4. The CNER issues its opinion and formulates its recommendations which it submits to the Minister with responsibility for Research and other Ministers concerned

Also, in 2000, Parliament tried to introduce new mechanisms in the management of public funds, following recommendations put forward by the OECD and by the Court of Auditors (Cour des Comptes). The outcome was the design of the Institutional Act of the Finance Law (Loi Organique Relative à la loi de Finances), which became operational the first time for the 2006 budget. This systematic tool is aimed at improving transparency of public sector accounts, improving the ability of the authorities to set spending priorities, and promoting a results based outlook (European Commission, 2006d). Performance of each program will be evaluated on the basis of three criteria:

1. Social and economic effectiveness
2. Quality of service
3. Efficiency
Practically, each program lists several specific results to which the program manager commits himself. It reports appropriations, main goals, performance indicators, expected results, and financial data. Public performance and efficiency will then be based on performance measurements. Together with this budgetary instrument, once put in place, AERES will have the duties to evaluate research programs.

Apart from these new tools, the Court of Auditors carries out auditing of public accounts. Its evaluation reports are about public bodies (but not only). It makes yearly public reports of which one focused on how resources made available by the previous year’s Finance Law have been used. It also produces reports as regards any organization of its choice. For instance, in 2005, it made available a report on the management of research at Universities.

As seen from the above, the French research system is undergoing many deep changes that predominantly affect the public research sector and to a lesser extent the private research sector. Since 1999 and the Law for Innovation and Research, successive Governments have shown their willingness to strongly modify the organization of research. The current objectives are to adapt the research system to the new challenges that have been identified by the Government in the Pact for Research. The main elements that are evolving are (European Commission, 2006d):

1. The mechanisms of funding of basic research. Traditionally, in France, public research is funded through contract mechanisms between the State and the research institutions such as universities and public research organizations (PROs). Along with direct funding, the Government has recently developed new instruments to fund research on the basis of projects whatever the institutional affiliation of the researchers. Currently, those competitive funds represent a marginal share of the government R&D spending. In 2005, the credit for payments of the ANR—€350 million ($461 million USD)—which funds research on the basis of calls for projects corresponded to only 2.5% of the GBAORD. However, this level is expected to increase significantly in the future. The scientific community is paying special attention regarding the
extent to which this increase will not be done at the detriment of funding provided to research structures.

2. **Human resource management.** Because of the ageing of population which affects research as well as the society as a whole, universities and PROs will have to recruit thousands of researchers in the coming years. These recruitments raise many problems such as the potential lack of researchers in some fields or the status of these researchers. Traditionally, French public researchers were given the status of ‘civil servant.’ The principle to pay researchers on the basis of the projects they carry out irrespective of their institutional affiliation is perceived as going against the principle of job security with which researchers were provided. This point is a major disagreement for those researchers gathered around the movement Let’s Save the Research.

3. **The role of universities in the public research system.** The French public system is dual in the sense that it is organized around universities and scientific umbrella organizations. The government has claimed its wish to reinforce the role of universities. At least two main issues may arise from this. The first is related to the universities themselves. Fundamentally, the basic traditional principle guiding the organization of universities in France responded until recently to a spatial planning logic. As a matter of fact, French universities are distributed everywhere in the French territory. The increasing aspiration to concentrate talents and resources to gain international visibility implies that this spatial planning principle may vanish. To support the gathering of public resources, two new instruments have been designed, namely the Research and Higher Education Clusters (PRES; Pôles de Recherche et d’Enseignement Supérieur) and the Thematic Advanced Research Networks (RTRA; Réseaux Thématiques de Recherche Avancée). In any case, this would mean that universities will most likely be given a larger autonomy than the one they actually have at the time being. The second issue is related to the reaction of the other Higher Education and Research actors such as the ‘grandes écoles’ and research institutes which were given an important role in the prevailing system and could feel threatened by this possible change of balance which would strengthen universities’ role in the national higher education and research system. In 2005, the General Inspection of the Administration of Education and Research (Inspection Générale de l'Administration de l'Education Nationale et de la Recherche) analyzed the regional strategies of
the main PROs. As far as the CNRS is concerned, the report underlined that it has difficulties to clearly identify the role it may play within the coming PRES.

4. **Private research.** Along research policies oriented towards the public sector, the government has increased its focus put on research in the private sector. The main idea is to incite companies to devote more resources to research. For instance, the Ministry Delegate of Higher Education and Research has designed specific measures such as the reform of the Research Tax Credit scheme (in 2004) or the status of ‘Young Innovative Company’ (in 2004). But above all, one of the main challenges of the French (public and private) research is related to the ability to reinforce the links between both. New schemes have been put in place to do so under the responsibility of the Ministry Delegate of Industry, like the Competitiveness Clusters or the Agency for Industrial Innovation (in 2005) which are aimed at providing support to private projects associating public partners.

In order to publicize its action regarding research policies, in 2005 the government designed a Pact for Research presenting the main challenges that the research system is presumably facing and the decisions to be taken consequently. According to the Pact, the main challenges that research policies have and will have to deal with in the near future are threefold (European Commission, 2006d):

1. **Reorganizing the public research system.** The Pact of Research implemented the wish voiced by the scientific community to reshape the role of the different research actors. However, the Law for Research voted in 2006 had left several blanks that would probably persist in raising questions in the future. According to the Estates-General of Research and Higher Education, the most important issues that remain opened are the attractiveness of research career and the associated question of the status of the researcher on the one hand and the orientation and steering of research policies at the Governmental level on the other hand. The scientific community is very concerned about the management of human resources in research. This point was raised during the consultation of researchers in the context of the Estates-General of Research and Higher Education organized in 2004. Their concerns are related to the ageing of the population and the disinterest of the young people for a research career. In order to identify the future needs of researchers, a ministerial
agency will be created in order to monitor employment and career of PhD holders both in the public sector and in the private sector. In spite of the reaffirmation of the predominant role of the Ministry Delegate of Higher Education and Research (which is currently under the aegis of the Ministry of Education, Higher Education and Research) in the definition and the implementation of research policies, in practice the Ministry does not have enough power to really lead the French research strategy. This was pointed out by the Court of Auditors (Cour des Comptes) in 2003 in its evaluation of research policies and was acknowledged by the Ministry of Research and New Technologies (now the Ministry Delegate of Higher Education and Research). As a matter of fact, the scientific community stressed, before the publication of the Project of Law, the need to create an independent Ministry of Research. In terms of institutional set up, in the last two decades, the status of the ministerial body in charge of research matters as well as its position in the Government has changed many times.

2. **Raise private investment in research.** In 2003, R&D expenditures performed by the business sector in France reached 1.37% of the GDP, which corresponded to a share in the total of GERD of 63%. However, the share of the GERD funded by domestic enterprises (52.1% in 2002) is in fact rather low in comparison with the main industrialized countries (in 2002 it was 65.5% in Germany, 73.9% in Japan or 64.4% in the USA, for example) and below the two-thirds target set in the Barcelona objectives.

3. **Reinforce the links between the public and the private sectors.** At the request of the President of the Republic, the CEO of Saint-Gobain, identified in 2004 some potential measures in order to renew the French industrial policy. His report assessed that the industrial policy should be defined around the State support of long term industrial technological programs. This would imply a stronger cooperation between private companies and public actors. The follow-up was the creation of the Agency for Industrial Innovation aimed at bringing together large firms, SMEs, and public research institutions.

According to von Tunzelmann and Mbula, (2003), “…there has been an increasing emphasis on evaluation of research during the last few years [in France, but]…this fact has not been necessarily reflected in the use of new evaluation techniques in general” (p. 17). Nevertheless, impact assessment and other alternative mechanisms of
evaluation have already been introduced in certain research institutions and prior to the creation of AERES, the Secretary General of CNER described the evaluation of research in France as "clashes of interest in the peer review system, too much promotion by seniority, and a lack of transparency in procedures" and that assessment structures are separated from decision structures (Berret, 2004, June).

Germany

In the following, Germany's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

As Europe's largest economy and second most populous nation, Germany remains a key member of the continent's economic, political, and defense organizations. European power struggles immersed Germany in two devastating World Wars in the first half of the 20th century and left the country occupied by the victorious Allied powers of the United States, United Kingdom, France, and the Soviet Union in 1945. With the advent of the Cold War, two German states were formed in 1949; the western Federal Republic of Germany (FRG) and the eastern German Democratic Republic (GDR). The democratic FRG embedded itself in key Western economic and security organizations, which eventually became the EU and NATO, while the Communist GDR was on the front line of the Soviet-led Warsaw Pact. The decline of the USSR and the end of the Cold War allowed for German unification in 1990. Since then, Germany has expended considerable funds to bring Eastern productivity and wages up to Western standards. In 1999, Germany and ten other EU countries introduced a common European exchange currency, the Euro. Germany had a population of 82.4 million in
Research Structure and Context

Most academic research in Germany is conducted either in organized research institutes or the scientific universities (Wissenschaftliche Hochschulen), which embody the long-standing "Humboldtian tradition of integrating teaching and research" (Geuna & Martin, 2001, p. 11). Research is also carried out in the polytechnics (Fachhochschulen), which are primarily teaching institutions. The higher education sector in Germany is made up of roughly two-hundred thirty full universities and universities of applied science (Fachhochschulen). The latter have the character of polytechnics in their close orientation on vocational courses and teaching over research (Orr & Pätzold, 2006).

Germany has three categories of public funding for university research (Orr, 2004a, 2004b). The first is institutional funding from the states (Bundesländer), which take the form of block grants to support basic infrastructure and staff. This funding accounts for nearly two-thirds of the total university expenditure (Kuhlmann, 2003) allocated by Bundesministerium für Bildung und Forschung (BMBF), which is responsible for direct federal contributions to public research institute sectors. The second funding source is capital grants provided jointly by the federal government (Bund) and the 16 states (Länder) for buildings and large-scale equipment. The third is third-party funds (Drittmittel), which are grants and contracts given by public institutions (Geuna & Martin, 2003). A large proportion of these funds are allocated by the German Research Foundation (DFG; Deutsche Forschungsgemeinschaft), jointly funded by the Bund and Länder, which are generally distributed to one or more of Germany's four largest research networks (OECD, 2003a): Max Planck Gesellschaft (MPG); Fraunhofer
Gesellschaft (FhG); Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF); and Wissenschaftsgemeinschaft Wilhelm-Gottfried-Leibniz (WGL).

Institutional and capital funds are allocated according to a profile that includes numbers of students, staff, and current spending. To determine research budgets, a 'R&D coefficient' is derived via surveys, to estimate "time spent on research and teaching" (Geuna & Martin, 2003, p. 286). Performance measures have not been used to allocate research funds, and there have not been evaluations for this purpose. This is partly because universities are mainly financed at the regional level, but it is also due to university hostility towards the competition that would be created. Among German academics, competition is not seen as a principle for advancing and encouraging research quality (Campbell & Felderer, 1997). Although there have been several evaluations of university research, these have not influenced funding (Daniel & Fisch, 1990).

Germany's institutions of higher education (HEI) are major contributors of R&D and increased their expenditure on R&D in the decade between 1992 and 2002 from €6.6 to €9.1 billion ($7.9 to $11.0 USD), an increase of 38% (Orr & Patzold, 2006). One third of this expenditure is financed through institutional funding by the individual Länder for 'their' HEIs. In principle, this is unspecific funding, which can be used by the institutions as they see fit. In most Länder a small proportion of this money is, however, provided on the basis of performance-indicators; for instance, number of doctorate graduates, third-party funding volume, and in two Länder, number of publications. However:

...regulations concerning the employment of academics do restrict capacity building strategies. According to their contracts, academic staff are allotted a normative amount of hours per week for teaching and for research. In the case of university professors the balance between these two activities is 50:50. This means that a university wishing to increase its research capacity by employing an additional professor will subsequently also increase its teaching capacity to the same proportion, and vice versa.
Two-thirds of R&D expenditure at HEIs is provided through third-party funding contracts from industry, government and the DFG. Many of the projects resulting from such funding contracts are realised through fixed-term contracts with research staff, but they are often supervised or led by professors. The Federal government and the Länder fund the DFG at a ratio of 58:42 (Waugaman, Friedrich, Tornatzky, & Schmidt, 2001) and nearly 40% of all third-party funding received by HEIs comes from the DFG (Orr & Pätzold, 2006). The largest proportion of DFG grants, approximately 60%, is for the general promotion of research and is allocated as a grant to individuals or a group of academics for a proposed project. In some cases the topics of these proposals emerge from scientists requesting support and in others the proposals are reactions to priority areas, which have been broadly defined by the DFG.

Evaluation Model

Attempts to rank individual and university research in Germany began in the mid 1970s and early 1980s (Alewell, 1990).272 One of the earliest efforts was WIBERA-Wirtschaftsberatung AG's273 publication of a list of research performance indicators in the early 1970s, which were then used for individual comparisons of research in English Studies and departmental comparisons in Law on the basis of publication counts and peer ratings (Daniel & Fisch, 1990). However, these efforts were almost completely abandoned in 1990 with the German reunification (Kehm, 1999).

In the late 1990s, a few Länder allocated additional resources on a competitive or performance-related basis, and since evaluations of research institutes have become more

---

272 Evaluation as a profession did not evolve in Germany until the mid 1990s (Struhkamp, 2005).
273 WIBERA-Wirtschaftsberatung AG is a mixed ownership (private and public) accounting and auditing company (Daniel & Fisch, 1990).
common with calls for evaluation of publicly-funded research (Campbell & Felderer, 1997). In 1994, the Bund and the sixteen Länder asked the Science Council (Wissenschaftsrat) to evaluate a group of eighty-two research and research-oriented institutions—the ‘Blue List’ institutes—over a period of five years, from 1995 through 1999 (Tegelbekkers, 1997). The Blue List institutes were evaluated on thirteen criteria of scientific quality and service, including for example, the number and quality of publications in national and international refereed journals and external funding.

Simultaneously, the Freie Universität in Berlin implemented its own evaluation mechanism, the results of which were used for the internal distribution of funding (Campbell & Felderer, 1997; Geuna & Martin, 2001, 2003). There have also been scattered evaluations of some of Germany’s scientifically oriented institutes, for example the MPG, which commenced in 1997. However, many universities, such as those in the Verbund Norddeutscher Universitäten (VNU) “perceive the use of research evaluation for internal funding allocation as retrospective in that it rewards those who have shown good performance in the past, while what is needed is a mechanism that helps universities to improve their performance” (Geuna & Martin, 2001, p. 12).

Systematic government evaluations of university professors or their research, however, was “unconstitutional” (Geuna & Martin, 2003, p. 287) under the Framework Act for Higher Education (Hochschulrahmengesetz). Under the Basic Law, the Hochschulrahmengesetz guaranteed the freedom of art and science, research and teaching. In research, this meant “that every scientific investigator is free, with regard to the choice of research topics, methodological principles and the evaluation and publication of his or her findings” (Fichtner, 1990, p. 332). However, in 1998, a major reform was made to the Hochschulrahmengesetz, with the Bundestag’s adoption of an

---

274 The name refers to the Blue Paper that was used for the initial list of institutes in the 1970s (Tegelbekkers, 1997).
amendment which made competition through “deregulation, performance orientation, and the creation of incentives” (Geuna & Martin, 2003, p. 287), which abolished the previous immunity of professors to external evaluation (ab Iorwerth, 2005). This made the evaluation of teaching and learning, and of research obligatory, which has, however, focused mostly on the former. Until recently, research evaluation has been as a subordinate objective. The reasons for this discrepancy are multifaceted:

Most reforms in the last decade have centred on teaching and learning and the challenge of expansion of student numbers, e.g. through the introduction of Bachelor and Master courses in line with the European-wide Bologna Process. Additionally, scientific research funding via third-parties (i.e. not institutional funding) is organised in a competitive system, so that evaluation may have been viewed in many states as less necessary (Orr & Pätzold, 2006, p. 4).

In 2000, the Monopolkommission (Orr, 2004b) published a report and the Wissenschaftsrat (2000) released Theses for the Future Development of the System for Higher Education and Research in Germany which argued that competition should become the leading principle for higher education policy. As yet, there are no large-scale federal systems for evaluating publicly-funded German university research. In part, this is explained by the fact that the universities are funded primarily by the Länder, but also in the widespread resistance to the concept of competition between universities (Geuna & Martin, 2001, 2003).

However, research evaluation in higher education has been conducted by the Academic and Research Commission (ARC) of the state of Lower Saxony. Within the New Public Management (NPM) program, ex-post assessment has developed at the institutional level and is used to inform, or even determine, public allocation of institutional budgets. NPM aims to afford institutions the greatest possible autonomy to make decisions on inputs and processes and judges only the merit of the outputs that result from these autonomous decisions (Orr & Pätzold, 2006). By this, it reforms an
academic tool, which is based on a low aggregation level (researcher, project), to an administrative tool, focused on a much higher level—groups of institutions. The higher up the level of aggregation a procedure is, the more important it becomes that the procedure is transparent and the evaluative criteria are standardised. For this reason, procedures used to allocate funding at institutional level are usually organised externally to the institutions being evaluated and based to a large extent on quantitative indicators (e.g. indicator-based funding).

The DFG uses a number of different formats for research evaluation. In line with its function as a funding body for research, the DFG has a long established procedure for the evaluation of proposals for funding, which is based on peer review of both past performance of academics and their institutions and the merit of their proposal. However, it has recently been recognised that the effectiveness of the projects and programs funded by the DFG should also be assessed.

Although organisations, which have received DFG funding, are obliged to write a concluding report on the implementation and results of their projects, this information has as yet not been analysed in any systematic way by the DFG. A new body set up at the start of 2006 and called the Institute for Research Information and Quality Assurance (IFQ) will take up this task. Additionally, at the end of the 1990s the DFG recognised a need to produce a cartographic analysis in the form of a ranking. This periodic ranking of research activities funded through public third-party funding analyses funding streams at institutional, regional and discipline level. In 2003 the report focussed on the network character of research and ranked inter alia research institutions according to the number of collaborative projects they were involved in.

A further development of DFG evaluative tasks is the implementation of the so-called Excellence Initiative, which the DFG jointly leads together with the Science
Council (WR; Wissenschaftsrat). This evaluation is essentially allocative. It is based on components of the DFG procedures, but extended to include institutions' strategic development plans. The Excellence Initiative will first be implemented in the autumn of 2006. Between 2006 and 2011 the program will allocate €1.9 billion ($2.5 billion USD) between both individual and collaborative programs from universities and non-university research organisations, based on their excellence in graduate (doctoral) teaching and world-class research. For those universities, which are successful in both areas, a third stream of funding to promote institutional strategies for top-level research will be available. It remains to be seen, what exact data and quality criteria will be used by the peers in the various assessment committees, although this is essentially a combination of existing procedures, all based on proposals written by applicant institutions. The WR is an agent of the state, funded by both the Federal Government and the Länder. It has the specific task of making recommendations on the structure and performance, planning and development of scientific institutions and this also involves cartographic evaluations based on institutional self-evaluations and peer review. However, these evaluations are carried out irregularly and are initiated on an ad-hoc basis by the state.

In 2000, each of the associations responsible for the non-university research institutes began to carry out its own research evaluations. These occur on a regular basis and are based in the main on peer reviews by academics from outside the respective associations, including foreigners. They are generally carried out both to inform decision-making within the associations and to report to external grant-givers on the performance of the associations and their respective institutes. For example, the Senate of the Leibniz Society, which has only external members, carries out a regular evaluation of the Leibniz Centers and uses these to make recommendations to the Federal government and the Länder on the funding of the centers and on their profile. At regional state level a
number of ad-hoc reports have been commissioned in the recent past to evaluate the
current structure of higher education performance (including research) and to make
recommendations for the future (e.g. in Hamburg, Schleswig-Holstein, North-Rhine
Westphalia and Bavaria). Besides these one-off exercises, Länder such as Baden-
Württemberg and Lower Saxony have introduced regional evaluation agencies. In Baden-
Württemberg evaluations are carried out by Evalag and encompass both research and
teaching. The studies compare discipline fields within different institutions of the same
type (i.e. universities), but rarely compare performance across the whole higher education
sector. The purpose is to inform both individual institutions and the state on
comparative performances (relative strengths and weaknesses).

In Lower Saxony two separate agencies were established for the evaluation of
teaching and research. Carrying out the procedure for research evaluation is one of the
main tasks of the Academic and Research Commission Lower Saxony. This institution is
unique in Germany, although the establishment of such a body was recently
recommended for Bavaria (Orr & Pätzold, 2006). Within the framework of quality
assessment, systematic, cross-regional and comparative research evaluation has been
carried out at universities and other non-university research institutions financed by
public funds since 1999. Its aim is to achieve valid statements regarding the quality of
research within institutions and subject areas within Lower Saxony and to summarise the
results cartographically.

In drawing up the basic structure of the procedure, the authorised bodies agreed
on a multidimensional, mainly qualitative procedure, to be applied in the form of peer
reviews within universities' subject areas and subject combinations as well as within other
HEIs. To this end, the ARC appoints panels of experts external to the Commission.
The procedure of informed peer review is run in accordance with internationally recognised standards, albeit with regional variation. A report from the institution to be assessed, portraying the achievements of the last five years as well as future plans and perspectives, is followed by a visit to the institution by the panel of experts. Talks thus take place at the institution between the HEI directors, the faculty directors and finally each individual research unit (these usually consist of single professorial chairs). The procedure can, therefore, be characterised as an in-depth evaluation from institutional down to individual research level.

The experts on the panel draw up a confidential draft assessment based on their impressions from the institution’s report and their site visit. The HEI assessed may then issue a statement on the draft via its directors. The experts may in turn respond to criticism or recommendations contained in the HEI’s statement. The ARC, which confers twice annually, is presented with the final report from the expert panel as well as the statements of the assessed institutions. It decides on a definitive version of the report, which is then in confidential form; that is, containing the names of individuals, is passed on to the Ministry and the directors of the assessed HEIs for their further use. A version of the report from which persons’ names have been removed is published and made generally accessible (e.g., on a dedicated Website).

The reports contain, on the one hand, qualitative assessments of the achievements of individual research units and of the subject area within the HEI as well as on a regional state level, and, on the other hand, the reports link these assessments to recommendations, equally related to these levels. These recommendations may touch upon such questions as the denominations of professorial chairs about to be vacated at research unit level; on the next level, the HEI may receive suggestions to restructure a subject area. From time to time, relocations from one site to another may be
recommended in order to reinforce the HEI's strengths, concentrate research potential and to raise achievement capacities and thus national as well as international prestige in the long term.

These sometimes extensive recommendations can, in turn, only be implemented at a variety of levels and, particularly if the recommendations take on greater proportions, through interaction with several agents. They require discourse and negotiation, not only within the HEIs themselves, but also between the HEIs and the Ministry. Instruments of implementation may include review committees and structural committees, for instance, coordinated by the ARC. In the last few years, the most important means of steering between the Ministry and HEIs have been jointly two-yearly target-agreements (Zielvereinbarungen). The ARC has established criteria "comparable to other national and international evaluations" on which the procedure is to be based. These criteria are intended to ensure recognition for the procedure and to allow a comparison of the results. On the one hand, it is a question of quality and relevance of the research results; on the other hand, effectiveness and efficiency within the research process take precedence. Each panel of experts appointed by the ARC is, therefore, required to answer the following questions, ranging from the level of individual research units to the subject area within the HEI and its achievement capacity on a regional state level:

What is the contribution of this research towards the prestige of the relevant discipline within the HEI as well as on a regional, national and international level?

Are funds implemented to achieve the intended effect whilst protecting standards of quality? How do these results compare to those of other locations?

Alongside the general criteria common to all procedures, particular significance is attributed to the definition of subject-specific criteria and the formulation of a subject-
specific research concept. The experts in the panel are at liberty to assign varying significance and chances of success to individual indicators in the relevant subject areas, particularly quantitative indicators such as third-party funding and international publications. The agreed assessment guidelines are thus applied uniformly across the regional state within the framework of procedures for the particular disciplines. The guidelines are also published at the beginning of every assessment report.

It is not expected that large-scale evaluation exercises like those conducted in the United Kingdom and Hong Kong will be installed in Germany in the near future (Kuhlmann, 2003). By 2004, it was announced that the Science Council had recommended a rating system much like the Hong Kong and United Kingdom RAE; the ratings are likely to based on a seven-point scale (a department would receive the highest mark only if more than half of its research activities are considered to be of top international quality; the lowest marks would go to departments whose output falls below national standards), conducted once every five or six years, and likely to affect future funding decisions, but they are not meant to become the decisive factor (Schiermeier, 2004). Moreover:

The Science Council is encouraging the development of both hard and soft indicators to be applied in varying degrees depending upon the field, which should spare anthropologists and quantum physicists from being measured on the same criteria (Schiermeier, 2004, p. 260).

As in most other industrialized countries, three layers of evaluation procedures in the area of research evaluation policy can be determined (European Commission, 2006e):

1. **First level—individual research performance.** The core was formed by peer reviews and later additional procedures to measure the research performance of individual researchers and groups (e.g., bibliometrics) as internal, scientific instruments for deciding on the allocation of promotional funds to research. Peer review procedures are in widespread use in the German research system, especially in the ex ante evaluation of projects in basic and long-term application-oriented research. Peer review is the
predominant evaluation instrument of the DFG. The DFG plays a central role in the promotion of basic research in universities, principally by granting individual researchers funds on application (the so-called standard procedure). Applications for grants are assessed by peers, who are elected every four years by the entire scientific community. Each expert is advised to judge the application on the basis of its scientific quality alone.

2. Second level—programs. Around a core of peer review procedures a 'shell' was formed which consists of impact analyses of R&D policy programs. The studies carried out in Germany mainly follow the tradition of the manifold work above all of American impact research, together with approaches from policy analysis, which has developed into an instrument of policy advice used in many fields. Impact analyses have gained acceptance in Germany since the 1970s in many political fields with the spread of program policy. Political claims for control call for an efficiency review and program evaluation and impact analysis have experienced a considerable upswing since then. The spread of evaluation is closely bound with the increase of strategic programs initiated by the EU Commission to promote R&D. As a rule, independent research institutes act as evaluators on behalf of R&D policy administrators. Since the mid 1990s many R&D policy programs have been launched as competitions, which aim to bring about a structural change in science and the economy: consortia of candidates (usually institutions) should in a self-organized process elaborate joint project plans and detailed goals. As a consequence, new evaluation designs are required. The experience with over two decades of program evaluation led to the establishment of a certain 'evaluation scene' in the German-speaking area, consisting of a group of experts and institutes from the field of economics and social sciences, who use a broad spectrum of concepts, methods and instruments and who have been organized professionally in a German Society for Evaluation (DeGEval) since 1998.

3. Third level—institutions. Here the performance of entire research institutions is dealt with. In Germany, the evaluations by the WR have been playing an important role for a long time (e.g., in the re-structuring of the 'research landscape' of Eastern Germany after reunification). Since the 1990s, evaluations of institutions have been carried out with greater frequency. In spring 1999 an international commission completed a system evaluation of the DFG and the MPG. At the same time, a system evaluation of the FhG was carried out. An evaluation of the institutions of the
WGL was completed in 2000; also a system evaluation of the national science centers in the HGF as well as a strategic evaluation of the joint industrial research mechanisms.

Hong Kong

In the following, Hong Kong's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

Occupied by the United Kingdom in 1841, Hong Kong was formally ceded by China the following year and various adjacent lands were added later in the 19th century. Pursuant to an agreement signed by China and the United Kingdom in 1984, Hong Kong became the Hong Kong Special Administrative Region of China in 1997. In this agreement, China promised that, under its 'one country, two systems' formula, that China's socialist economic system will not be imposed on Hong Kong. In 2006, Hong Kong had a population of 6.9 million (Central Intelligence Agency, 2006).

Research Structure and Context

Research in Hong Kong is funded by a dual-support system through the University Grants Committee (UGC) and the Research Grants Council (RGC). The UGC is a non-statutory body that acts as an advisor to the government of Hong Kong, while the RGC, under the aegis of the UGC, allocates project grants (French, Massy, & Young, 2001). Hong Kong's UGC was established in 1965, and was responsible for advising the government on the development and funding of the then two institutions of higher education, namely the University of Hong Kong and the Chinese University of Hong Kong. The UCG came into being during the 1964 Budget Debate as a result of
members of the Legislative Council arguing that a committee similar to the British University Grants Committee (BUGC) should be set up in Hong Kong to advise the government on the facilities, development, and financial needs of the universities. The UGC was formally appointed in October 1965, with principles and practices based on the British model.

The UGC was renamed the University and Polytechnic Grants Committee (UPGC), in 1972 to reflect the inclusion of the then Hong Kong Polytechnic (now The Hong Kong Polytechnic University). In 1983, the former Hong Kong Baptist College (now Hong Kong Baptist University) was brought within the ambit of the UGC, followed the next year by the then City Polytechnic of Hong Kong (now City University of Hong Kong) and in 1991, the Hong Kong University of Science and Technology and a former post-secondary college, Lingnan College (now Lingnan University) were added. Following the adoption of university titles by the two polytechnics and the Hong Kong Baptist College, the UPGC reverted to its previous title of University Grants Committee in November 1994 (University Grants Committee, 2005).

At present, the UGC funds City University of Hong Kong (CityU), Hong Kong Baptist University (HKBU), Lingnan University (LU), Chinese University of Hong Kong (CUHK), Hong Kong Institute of Education (HKIEd), Hong Kong Polytechnic University (PolyU), Hong Kong University of Science & Technology (HKUST), and the University of Hong Kong (HKU). All of these institutions are “statutorily autonomous corporations, each with its own ordinance and governing council, enjoying academic freedom and very considerable institutional autonomy, subject to the constraints of financial dependence and public accountability” (French, Massy, & Young, 2001, p. 35).
Evaluation Model

In 1991, the UGC determined that its methodology for the assessment of the institutions' recurrent funding requirements should be refined, particularly the provision for research funding (French, Massy, & Young, 2001). During the period from 1991 to 1994, the UGC implemented a system of research assessment modeled closely after the United Kingdom's RAE of 1992 (Davies, 1994). The aim of the assessment was to inform the distribution of the UGC research fund ('R' funding) within recurrent block grants and to “discharge public accountability and to induce improvements in research” (University Grants Committee, 2004, p. 1). These assessment exercises were to take place in intervals that coincided with the UGC's triennial funding cycle.

Hong Kong's first institutional research assessment took place in 1993, implemented with the help of external consultants from the United Kingdom, and was followed by exercises in 1996 and 1999. Subsequent to the completion of the 1999 REA, the UGC decided that future exercises should be undertaken at six-year intervals, with the next to take place in 2006 (University Grants Committee, 2004).

The first Hong Kong RAE, like the United Kingdom's, was intended to rate cost centers (i.e., departments) and institutions, not individual researchers. However, unlike the United Kingdom, the Hong Kong exercise was conducted solely on the basis of the number of researchers in each cost center. Cost center's submissions were assessed by eight panels and the results were aggregated, with subject-specific weightings, over all cost centers in each institution to determine each institution's R-allocation (French, Massy, & Young, 2001). Each RAE panel consisted primarily of local academics from UGC-funded institutions together with one United Kingdom expert. The key task for

275 The UGC assesses institution's triennial recurrent funding needs separately for research (R) and teaching (T).
each panel was to determine the number of active researchers out of the nominated researchers from each cost center, on the basis of the quality of the best three research outputs from each nominated researcher (University Grants Committee, 1996). Thus, the general scheme of the first, and subsequent, Hong Kong RAEs was to determine the percentage of full-time equivalent active researchers in each cost center whose research was judged to have reached or surpassed the UGC quality threshold. The quality threshold standard was defined as “quality of output equates to an attainable level of excellence appropriate to the discipline in Hong Kong, and showing some evidence of international excellence” (University Grants Committee, 1996). The percentage, $p$, for allocating R funding was determined by the formula as given in Equation 2 (University Grants Committee, 1996),

$$p = 100\% \times \frac{A}{T}$$

where $T =$ the total number of academic staff in the cost center who meet the criteria regardless of the source of funding and whether they submit research output items for assessment and $A =$ the total number among these who are judged by the Panel to be research active, including fractional counts.

In contrast, the United Kingdom’s RAE assigns quality ratings to the submissions of staff nominated by institutions as research active and the quality ratings are then multiplied by the number of nominations and field-specific units of funding as the basis for block grant allocation (Geuna & Martin, 2003).

The UGC introduced numerous refinements for the second, 1996 RAE, including the introduction of larger assessment panels, greater numbers of overseas reviewers, and clarification of eligibility and purposes of the assessment exercise (French, Massy, & Young, 2001).
Following extensive consultation within the higher education community and more widely, significant changes were introduced for the third Hong Kong RAE (French, Massy, & Young, 2001; University Grant Committee, 1999a), conducted in 1999. Among these was the UGC’s adoption of the Carnegie Foundation’s definition of research and research-related scholarship (Boyer, 1990), which are:

1. **Discovery.** Contributes not only to the stock of human knowledge but also to the intellectual climate of an institution

2. **Integration.** Work that seeks to interpret, draw together and bring new insights to bear on original research

3. **Application.** Creating new intellectual understandings arising out of theory and practice

4. **Teaching.** Transforms and extends knowledge while transmitting an intelligible account of knowledge to the learners

Other changes introduced in 1999 included, but were not limited to, dropping the term ‘active researcher’ to clarify that the exercise was not intended as an assessment of individual researchers, asking institutions to map their research strategies in order to provide contextual information to be considered in addition to numeric indices, and dissemination of panel operating guidelines describing the standards and criteria to be used in the assessment for greater transparency (University Grant Committee, 1999a). In addition, the number of panels was increased to twelve, as opposed to the eight used in the 1993 and 1996 exercises. These panels also included a wider and more international membership. In all, there were one-hundred eighty panel members, of whom forty-two were from outside Hong Kong (French, Massy, & Young, 2001); more than one-third (39%) of whom had experience in the previous RAE. In contrast, the 1996 exercise one-hundred eleven panel members, of whom fifteen were from outside Hong Kong (University Grants Committee, 2004).
The eight UCG-funded institutions were to submit, for each eligible academic staff member,²⁷⁶ up to five research outputs within the assessment period (1995-1998) for assessment. Research outputs were defined as (University Grants Committee, 1999a):

1. Any publication, patent, or other artifact, provided that it was:
   a. Published or made publicly available within the assessment period, or
   b. Not yet published, but officially accepted for publication in that period

2. Other output that may or may not be in publishable form (e.g., drama, concert performance, video tape, computer software program, buildings, or creative work that could be evaluated for merit and an assessment obtained)

Furthermore, the institutions could make separate submissions of one exceptional research output which did not fall within the assessment period, for one individual staff member. While a maximum of five research outputs per eligible staff member could be submitted, the UGC stresses that panels should examine only the three judged as being of the highest quality and ignore the remainder. Submissions were judged using the ‘scoring schema’ (University Grants Committee, 1999b) presented below in Table 4. These scores, or grades, were intended to allow the UGC to view the quality, rather than quantity, of Hong Kong research across different categories of scholarship, although they were not a factor for determining the \( p \)-index for allocating R funds.

²⁷⁶ Eligible staff were defined as those who held a paid appointment for a continuous period of twenty-four months or more covering the specified census date, whether or not the continuous appointment was principally before or after the census date, and not also holding a paid position at another institution (French, Massey, & Young, 2001; University Grants Committee, 1999a).
Table 4
Hong Kong's Research Assessment Exercise Scoring Schema

<table>
<thead>
<tr>
<th>Scoring schema</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>Masterpiece, cannot be ignored by anyone working in the field. Single item over 3-4 years is already good performance.</td>
</tr>
<tr>
<td>A</td>
<td>Highly innovative and significant. Probably noticed by anyone working in the broad field.</td>
</tr>
<tr>
<td>B</td>
<td>Innovative and significant, makes a valuable contribution to the field. Meets attainable standard of excellence common in the mainstream of the field.</td>
</tr>
<tr>
<td>C</td>
<td>A useful contribution, but possibly short of the attainable standard of excellence common in the mainstream of the field.</td>
</tr>
<tr>
<td>D</td>
<td>Standard below C.</td>
</tr>
</tbody>
</table>

Almost 19,000 research outputs from more than 4,200 eligible staff from the eight UGC-funded institutions were submitted for assessment in 1999 (University Grants Committee, 1999b). Each of the items submitted were classified, by the eligible staff person or institution, into one of the four Carnegie scholarship categories ('discovery', 'application', 'integration', and 'teaching'). Most were classified in the scholarship of 'discovery' category, although considerable numbers of submissions were classified in the 'integration' and 'application' categories (French, Massy, & Young, 2001; University Grants Committee, 1999b). Although the panels were instructed to grade the quality of research outputs using the UGC scoring schema, little direction was given for defining quality in terms of the four scholarship categories. The only guidance for rating submissions on the scoring schema was the UGC's Guidelines for Panel Members (1999c), which stated:

...the quality of each item should be judged on its own merits and not solely in terms of its category (e.g., a journal paper is not necessarily of higher or lower merit than a book chapter, nor is a refereed article necessarily of higher or lower merit than an unrefereed one), venue or language of publication. Panels should recognize that there could be quality output items in venues that may not be prestigious. In these cases,
and in any case when in doubt, the Panel (or designated member(s)) should study the item in question and not judge it automatically according to the venue (University Grants Committee, 1999c).

In response to concerns raised across "various sectors of the local higher education community regarding the existing RAE mechanism" (University Grants Committee, 2004, p. 4) the UGC formed a Research Ad Hoc Group (RAG) to make recommendations for improving the 2006 and future exercises. Based on RAG findings and recommendations and a number of modifications were made for the forthcoming 2006 RAE. Among these was the allocation of a small percentage of total R funding of the block grant to recognize and reward research performance at the top end, increasing the number of allowable submissions from five to six, and panels were to judge the best four submissions rather than three. Furthermore, although the 2006 RAE will maintain the Carnegie definition of research and the four research-related scholarships (Boyer, 1990), a second Carnegie Foundation study titled Scholarship Assessed: Evaluation of the Professoriate (Glassick, Huber, & Maeroff, 1997) will be applied to "ensure that scholarly work in areas both within and outside discovery can be appropriately recognized and rewarded" (University Grants Committee, 2004, p. 22).

The intention of the aforementioned modification is to evaluate Hong Kong research outputs according to a common set of criteria, referred to as 'quality standards of excellence' (University Grants Committee, 2004, p. 22) and serve as the basis for the RAE's 'quality threshold'. Additionally, to reduce divergence in panel judgments on the basis of the quality threshold, the UGCs introduced the following definitions for panels to apply in assessing the quality of submitted research outputs in 2006 (University Grants Committee, 2004, p. 17):

1. **International excellence.** This should not be equated with output

---

277 Defined as "an attainable level of excellence appropriate to the discipline in Hong Kong, and showing some evidence of international excellence" (University Grants Committee, 2006, p. 16).
items published outside of Hong Kong, or the region; rather it is intended that evaluation should be made with reference to the best international norms in the mainstream of that discipline or sub-discipline. It is possible that in some particular disciplines, such norms are set by output items published in Hong Kong or the region.

2. *International versus local.* A distinction should be made between (a) a publication that is local because it addresses local issues, and (b) a publication that is local because it does not meet the standards of rigor and scholarship expected internationally in the mainstream of that discipline. In the former case, the item will not be discounted; in the latter, it will be.

Overall, the administrative and compliance costs of implementing Hong Kong's RAE is less than 1% of the funding distributed under the system (Web Research, 2004).

Hungary

In the following, Hungary’s research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

**Historical Overview**

Hungary was part of the polyglot Austro-Hungarian Empire, which collapsed during World War I. The country fell under Communist rule following World War II. In 1956, a revolt and announced withdrawal from the Warsaw Pact were met with a massive military intervention by Moscow. Under the leadership of Janos Kadar in 1968, Hungary began liberalizing its economy, introducing so-called ‘Goulash Communism.’ Hungary held its first multiparty elections in 1990 and initiated a free market economy. It joined NATO in 1999 and the EU in 2004. Hungary had a population of 9.9 million in 2006 (Central Intelligence Agency, 2006).
Research Structure and Context

There are three major sources of R&D funding in Hungary: the central budget (€380 million [$505 million USD]); businesses (€270 million [$359 million USD]); and foreign funding (€75 million [$99 million USD]), in 2004. The bulk of domestic public funding goes to R&D institutes (€185 million [$286 million USD]) and the higher education sector (€144 million [$191 million USD]). The decisive source of funding for these R&D performers, in turn, is the state budget at 86% and 81%, respectively (European Commission, 2006f).

Firms’ R&D units constitute the largest research performing sector, and 77% of funding is provided by businesses themselves. The overwhelming majority of foreign funding, almost 73%, is obtained by this sector, while the share of state funding is a mere 4%.

There are five main channels of public funding for RTDI activities conducted in Hungary. In several cases these channels are actually allocating ‘mixed,’ public and private, funds. Three of these channels provide institutional, or core funding (European Commission, 2006f):

1. Research activities conducted at universities are financed by the normative research support, that is, institutional funding, as well as by various governmental funds and programs (e.g., the Higher Education Research Fund), offering competitive grants. The total public R&D funding for the higher education sector was around €144 million ($191 million USD) in 2004.

2. The institutes of the Hungarian Academy of Sciences (MTA) are financed by the central budget, distributed by the headquarters of MTA, as well by funds raised by applying for domestic and international grants. The total budget of the MTA was €122 million ($163 million USD) in 2005. The MTA also supervises the activities of the Hungarian Scientific Research Fund, supporting basic research projects, young researchers’ projects, and R&D infrastructure development on a project base, as well as
the Bolyai Janos Research Scholarship scheme, funding for outstanding researchers in the form of a stipend. These two schemes allocate competitive grants.

3. Several ministries provide funding for R&D activities in various ways: running their own R&D institutes, offering a mix of core funding and competitive grants for them, or only providing competitive grants to R&D units, regardless of their owners. In total, less than €24 million ($31 million USD) was spent on R&D from these sources in 2005.

The remaining two channels only offer competitive grants (European Commission, 2006f):

1. The Research and Technological Innovation Fund provides support for all sorts of RTDI performers (public, private, non-profit, and their consortia). It is financed by an innovation contribution to be paid by firms, and a matching fund from the central budget. Strategic decisions on the use of the Fund are made by the Research and Technological Innovation Council: which sorts of technology policy schemes to be launched, and how much funding to be allocated to the specific schemes. These decisions are prepared by the National Office for Research and Technology, and then implemented together with the Agency for Research Fund Management and Research Exploitation.

2. The Economic Competitiveness Operative Program (ECOP) of the first National Development Plan (2004-2006) funds six RTDI-related measures (spending 22% of the total ECOP budget), relying on EU Structural Funds (SF) and national co-financing. The budget for these RTDI schemes is approximately €140 million ($186 million USD) for the period of 2004-2006, of which approximately €100 million ($133 million USD) is financed by the EU SF.
Evaluation Model

Throughout Eastern Europe after 1945, national research systems were based on the Soviet tripartite model in which universities focused on teaching, basic research was conducted in Academy of Sciences’ institutes, and applied research was carried out in institutes under the various ministries (Frankel & Cave, 1997). Governments also adopted the Soviet system for funding research largely through block grants to institutes, a system in which the scientist-administrator had great power over the internal distribution of funds. Favoritism and political connections often gave rise to poor quality researchers and research teams with the right connections being funded, while high-quality researchers and research teams were under-funded (Geuna & Martin, 2003).

This system has changed radically over the last ten years as Eastern European countries have been transformed from centrally-planned to more open, competitive market economies. This has had a major impact on science policy. The autonomy of science in terms of self-evaluation through peer review, an autonomy completely subordinated to the central plan during the Communist era, has been restored. In the early part of this transition, economic crises saw research evaluation emerge as an important tool mainly to examine where to cut budgets without completely destroying research activities (Geuna & Martin, 2003). Peer review has now become the main evaluation mechanism used to allocate research funds (Hangos, 1997; Zilahy & Láng, 1997).

For several decades, MTA has conducted a fairly comprehensive evaluation of all its institutes almost every year. The evaluation in 1992 had a particular impact as it coincided with the cutting of research funds as a result of economic crisis. The evaluation was conducted in two phases. In the first, each institute was evaluated with a
view to restructuring the Academy’s research network and reformulating its tasks. In the second phase, each research unit or group within the institutes was evaluated to provide the basis for a more differentiated distribution of funds (Geuna & Martin, 2003). The evaluation method was peer review combined with quantitative, bibliometric indicators (Vinkler, 1997). The findings led to recommendations covering the function of the Academy’s research network, the management of human resources, financial conditions, and organizational changes (Zilahy & Láng, 1997).

More recently, a visible impact of EU practices can be detected, however, in the documents of policy schemes operated since 2004; in most cases, indicators are specified ex ante for the measurement of their results/impacts. A monitoring strategy is also being devised, based on the following underlying principles: policy-relevant programs and projects (e.g., those schemes and projects where a considerable amount of money is spent, or those pursuing essential policy goals) would be thoroughly monitored, while those with less significant funding (e.g., small grants for international project preparation) would be checked only by financial and administrative criteria (European Commission, 2006£).

Ex ante evaluations of the research policy measures launched in 2004 as part of the Community Support Framework—of which six are directly relevant for RTDI and a further two are of indirect relevance—had to be carried out, as requested by EU rules. Internal, self-evaluation of policy measures, mainly conducted by those government officials who designed the measures themselves, are carried out whenever a decision is due concerning the renewal of a given measure. The results of these internal self-evaluation exercises are not published.

External evaluations of policy programs have only been carried out occasionally in recent years; although, there had been conscious efforts to introduce systematic
evaluations in the mid-1990s. More recently, two research policy program evaluation projects have been completed. One concerns the Hungarian Technology Foresight Program (TEP), conducted by an international panel of experts. Although hard copies of TEP reports were published in 2001, and in the same year, the European Commission’s Joint Research Center (DG JRC) offered technical and financial assistance to carry out the evaluation, which started in 2003 without external funding. The evaluation report was discussed in 2004, and then published on the Internet. No action has been taken so far, though, in reaction to the evaluation.

The second example is the evaluation of a policy scheme to promote academia-industry co-operation, called Co-operative Research Centers. The evaluation report (in Hungarian) has been made available electronically. A major development in this respect is that the Law on Research and Technological Innovation of 2004, which stipulates that policy programs must be evaluated. Yet, besides the one mentioned above, not a single research policy program has been evaluated (European Commission, 2006).

The fundamental challenge for Hungary is to significantly enhance its international competitiveness and then maintain it. Macroeconomic pressures, notably budget, trade, and balance of payment deficits, also call for a successful, competitive economy, supported by a strong national innovation system, both in terms of its elements and the communication and co-operation among the various players. Clearly, a well-functioning newly independent state (NIS) requires adequate human resources, too (in terms of quantity and quality). Yet, only a small proportion of young talents opt for STI careers, while experienced researchers leave Hungary, or swap for better paid, more prestigious jobs. These trends can only be reversed, or at least slowed down, by offering attractive conditions for scientists and research engineers. There is a significant gap in terms of human resources for R&D and innovation, too. Further, brain drain is a serious
threat, which is harmful both from an economic and a social point of view. The exploitation and commercialisation of R&D results is not sufficiently fast and widespread, partly due to the fact that academia-industry co-operation is weak. As noted by the European Commission’s ERAWATCH, there are a number of symptoms of the broad challenges mentioned above, facing Hungary, as reflected in various RTDI indicators and Hungarian policy (European Commission, 2006f):

1. Business expenditures on R&D are at 0.33% of GDP (in 2004), that is, less than one third of the EU average.

2. The total Hungarian R&D expenditures were 0.88% of the GDP in 2004.

3. Academia-industry co-operation is insufficient. Recently, a number of schemes have been introduced to promote academia-industry co-operation.

4. The bulk of indigenous SMEs struggle for day-to-day survival, and are not engaged in innovation activities. Indigenous SMEs often lack sufficient financial resources and managerial skills to engage in RTDI activities, and join international production and innovation networks.

5. The ratio of science and engineering graduates among people aged between 20 and 29 was 4.8% in 2004, which leaves Hungary in the 21st position in the EU. Further, Hungary under-performs in terms of the share of the working age population with third-level education.

6. Policy-making processes are not sufficiently transparent due to the lack of meaningful dialogues with stakeholders and experts.

7. There is a strong tendency to ‘reduce’ RTDI into research in advanced scientific fields, ‘equate’ R&D with innovation, and neglect the variety of types and sources of knowledge required for successful innovation processes.

8. Public support to RTDI is not efficient and effective because of the lack of policy co-ordination.

9. Modern policy-making methods are rarely used, although suggested by the Science and Technology Policy,
Competitiveness Advisory Board and the first Hungarian TEP (1997-2000). No policy reviews have been produced so far, in Hungary, nor has a systematic international comparative policy analysis been used to assess STI policies. The application of methods preparing policy decisions, however, has not been included in the Law on Research and Technological Innovation—although, suggested by independent experts on several occasions when the draft legislation had been discussed. Evaluation of STI policy measures has become compulsory since 2005—but, only one policy program has been evaluated so far. Thus, it cannot be established if public money is spent in an effective and efficient way, to achieve the desired objectives.

10. Policy schemes are changed too frequently and similar or the same objectives are supported by several schemes. This leads to increased search and administrative costs for potential research applicants.

Ireland

In the following, Ireland's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

Celtic tribes arrived on the island now known as Ireland between 600-150 BC. Invasions by Norsemen that began in the late 8th century were finally ended when King Brian Boru defeated the Danes in 1014. English invasions began in the 12th century and set off more than seven centuries of Anglo-Irish struggle marked by fierce rebellions and harsh repressions. A failed 1916 Easter Monday Rebellion touched off several years of guerrilla warfare that in 1921 resulted in independence from the United Kingdom for twenty-six southern counties; six northern counties remained part of the United Kingdom. In 1948 Ireland withdrew from the British Commonwealth and in 1973 it joined the European Community. Irish governments have sought the peaceful
unification of Ireland and have cooperated with Britain against terrorist groups. A peace settlement for Northern Ireland, known as the Good Friday Agreement and approved in 1998, is being implemented with some difficulties. In 2006, Ireland had a population of 4.0 million (Central Intelligence Agency, 2006).

Research Structure and Context

The Irish public research funding system is comprised of a number of funding bodies, each of which report to individual government departments. The Office of Science and Technology (OST) which is located within the Department of Enterprise, Trade and Employment has responsibility for the overall national science budget. The two main research funding organizations are (European Commission, 2006g):

1. The Science Foundation Ireland (SFI), which was established in 2000 to administer Ireland’s technology foresight fund of €646 million ($859 million USD) to support academic researchers and research teams in the fields underpinning two broad areas, biotechnology and information communications technology. It operates under the auspices of the Department of Enterprise, Trade and Employment.

2. The Higher Education Authority (HEA), which is the funder of the HEA Block Grant provides research funding in the third-level sector and the Program for Research in Third Level Institutions (PRTLI) which provides financial support for institutional strategies, inter-institutional collaboration, large-scale research programs, and research infrastructure. A total of €605 million ($805 million USD) has been provided to the higher education sector under the PRTLI during the period 1999-2006; a substantial amount of this funding was provided from private philanthropic sources.

Other funding organizations include the Irish Research Council for Science, Engineering and Technology whose Embark Initiative provides funding for postgraduate research students and sectorally-focused research funding and performing agencies such as the Marine Institute, the Health Research Board, the Environmental
Protection Agency and Teagasc (the agricultural research body). The most recent available data shows that the total State funding of S&T activities from government departments, agencies, and offices increased by 9.7% between 2003 and 2004, from €1.88 billion ($2.50 billion USD) to €2.06 billion ($2.74 billion USD). Total state funding of S&T includes expenditure by the exchequer, expenditure by the EU, and finally receipts from the earned income of activities. State funding is estimated to have risen by a further 6.5% in 2005 to total €2.20 billion ($2.97 billion USD).

The Irish government currently invests in a wide range of R&D-based programs, including (European Commission, 2006g):

1. Programs to support research and development performed in the higher education sector (e.g. funding given via SFI, the HEA’s Program for Research in Third Level Institutions, the Irish Research Council for Science, Engineering and Technology, and the Irish Research Council for the Humanities and Social Research)

2. R&D programs performed in the government sector by government departments, agencies, Teagasc, and the Health Research Board

3. Projects to assist businesses develop and increase R&D capabilities via programs operated by Industrial Development Agency (IDA) Ireland (the agency responsible for promoting inward investment) and Enterprise Ireland (the agency charged with the development of indigenous manufacturing and internationally traded service sectors)

4. Supporting R&D infrastructure across all sectors of performance

Total government funding (including direct and indirect sources) accounted for 83% of all research income in the higher education sector in 2004, increasing its funding share from the 79% in 2002. Other sources of research income for the higher education sector include funding from the EU, foreign sources, Irish enterprises and other national funding (including internal funds).
Seven universities remain the dominant performers of R&D across the higher education sector and continue to account for the majority of HERD (94%). However, there was also a strong increase in the amount of expenditure dedicated to R&D activities across the fourteen institutes of technology between 2002 and 2004. An intensive technology foresight program carried out in 1998-1999 involving representatives of the public sector, academia, and industry concluded that biotechnology and information communications technology had the potential to be important engines for future growth and Ireland should develop a world class research capability in these disciplines as an essential foundation to capitalize on that growth. The Irish Council for Science, Technology and Innovation (ICSTI) which coordinated the technology foresight program specifically asked the government to establish a fund which would enable Ireland to become a center for world class research excellence in niche areas of information communication technology, biotechnology, and their underlying sciences. ICSTI argued that without such a research capability to support the technology-based industries, which now accounted for more than two thirds of manufacturing output in Ireland, it would be impossible to sustain the momentum built up by the inward investment policy (European Commission, 2006g).

The Irish Government then decided to commit €646 million ($858 million USD) to a Technology Foresight Fund for academic researchers and research teams in biotechnology and information communication technology. SFI was to administer this investment fund. Through its investments in biotechnology and information communication technology research, SFI is seeking to support knowledge creation and human capital development.

The Enterprise Strategy Group, which was established in 2003 to prepare a report that would serve as a blueprint for an enterprise strategy for growth and
employment in Ireland, recommended that the government increase the level of public funding for applied research and in-firm R&D should be progressively increased to match that invested by the Department of Enterprise, Trade and Employment in basic research. This includes support for in-firm capability development, commercialization, cluster-led academic research and innovation partnerships (European Commission, 2006g).

In total, there are 13 research funding agencies in Ireland. In terms of subject breakdown, the Research Councils in Ireland are divided into Humanities and the Social Sciences (IRCHSS), Science, Engineering and Technology (IRCSET), and Health (HRB), plus some smaller institutes. The first has recently launched a Project Funding Scheme to support team-based research in relation to economic, social, and cultural development. The second launched its first program, the Basic Research Grant Scheme, jointly with Enterprise Ireland, in 2001. SFI was founded as a result of a 1998 Foresight Exercise in 2000 and currently aims at recruiting and retaining research groups and centers, with biotechnology and information technology identified as the main targets. Much emphasis is placed on international as well as national peer review, and on European projects (von Tunzelmann & Mbula, 2003).

**Evaluation Model**

Policy and program evaluations have become an important element of the Irish policy development and review infrastructure. A major contributing factor here has been the influence of EU practices in relation to the evaluation of policies and programs. As such, all of the major research policies and programs in Ireland are now subject to regular review. The National Development Plan 2000-2006 (NDP) and the Community Support Framework have been subjected to an ex ante review and, more recently, a mid-
term or interim review. The results of these evaluations and other inputs are important sources of data for policy-makers in framing the successor plan to the current NDP. For example, one of the recommendations of the mid-term evaluation report of the NDP was that because of the importance of the Lisbon strategy additional funds should be allocated to RTDI within the Productive Sector Operational Program for the remaining period of the NDP. When the current NDP is completed a final evaluation will be undertaken (European Commission, 2006g).

Evaluations are also undertaken at an organizational and program level. The government requested Forfás, the national policy and advisory board for enterprise, trade, science, technology and innovation, to undertake an evaluation of the performance and impact of SFI since its creation. SFI was set up by the Irish government to undertake and support strategic research of world class status in key areas of scientific endeavor which would underpin economic development. It was designed to be a key mechanism in the rapid evolution of Ireland to a knowledge society (Department for Trade, Enterprise and Employment [DETE], 2006; Inter Departmental Committee on Science, Technology and Innovation [IDCSTI], 2004). The government identified three broad areas in which it wished the evaluation to focus (European Commission, 2006g):

1. Appropriateness or efficacy
2. Effectiveness
3. Efficiency

The Lisbon Strategy is intended to deal with the low productivity and stagnation of economic growth in the EU, through the formulation of various policy initiatives to be taken by all EU member states. The broader objectives set out by the Lisbon Strategy are to be attained by 2010. It was adopted for a ten-year period in 2000 in Lisbon, Portugal by the European Council. It broadly aims to make Europe, by 2010, the most competitive and the most dynamic knowledge-based economy in the world.
Forfás gathered an international panel of experts under the chairmanship of Richard Brooks, Director of the Leverhulme Trust in the UK, and the panel made a number of recommendations in their 2005 report which have been endorsed by the government. In addition, major program evaluations include the evaluation of the Program for Research in Third Level Colleges. This evaluation was also carried out by a team of international experts at the request of the HEA, which administers the program on behalf of the Department of Education and Science. The Assessment Committee’s report made a number of recommendations directed at the government, the HEA, and the third-level institutions.

Within Forfás, the Evaluations Department within the Science, Technology and Innovation Policy and Awareness Directorate, has been established to evaluate science and technology programs on behalf of DETE. The Evaluations Department, for example, provided support services to the international panel of experts which carried out the evaluation of SFI. Furthermore, the Office of the Chief Science Adviser includes as one of its six key functions the overseeing of a system of independent evaluations of science, technology and innovation policy and programs, with particular reference to cross-cutting issues (European Commission, 2006g).

Ireland has gone through dramatic changes in public policy towards science, technology and innovation (ST&I) in recent years. The launch of the NDP 2000-2006 represents a strong commitment of the Irish government to scientific and research activities. A reason for these changes can be found in the that, although the public funding for ST&I has approached international norms, Ireland still lags behind OECD countries in terms of evaluating the benefits of such funding provision (Department for Trade, Enterprise and Employment, 2006; Inter Departmental Committee on Science, Technology and Innovation, 2004). A scarcity of expertise has been identified in terms of
evaluating the outcomes of expenditure on ST&I within the Irish public sector bodies (von Tunzelmann & Mbula, 2003).

ICSTI has recently published a number of reports, including the report of a task force on the Embark Initiative. This report examined the multiple evaluation practices in different organizations relating to STI. Examples of units with evaluation units include: the HEA (e.g., collection of all personnel data, ad hoc evaluations of all programs), the universities with a QAQI system, Teagasc, and Enterprise Ireland (evaluating all their technological programs).

The council concluded that inputs are quite well monitored in Irish research evaluations, but not their outputs or impacts. To account for this, the report begins by contrasting the traditional linear model approach, also implied in neoclassical economics approaches, with the broader evolutionary/institutional model operating typically through systems of innovation, allowing for multiple feedbacks and interactions (von Tunzelmann & Mbula, 2003). In terms of measurement, the former leads to an input-output approach whereas the latter leads more naturally to a throughput approach in which the focus lies more on the process of research rather than the products. Frequently, though not necessarily, the former measures are used in a summative fashion to inform a particular policy decision (like a United Kingdom RAE rating), while the latter are used in a formative fashion to improve research performance.

Principal indicators and evaluation methods in use tend to be implicitly related to the former input-output approach, at various levels (from the macro to the micro), which include: R&D expenditures, human capital indicators, bibliometric analyses, patent analyses, technological intensities of production, technological trade indicators (e.g., technology balance of payments, technology levels of exports), growth accounting analyses, measures of price, and quality changes (European Commission, 2006g).
Furthermore, the ICSTI reports suggested techniques suggested that might be used in throughput analyses, such as peer review, surveys/interviews, case studies, cost-benefit analyses, productivity spillover models, knowledge flow models (linkages), and technology foresight exercises. However, the ways in which these are generally implemented "fall some way short of theoretical desiderata, and the resource cost of doing them well could be extremely high. On the other hand, the throughput approach is much more in line with modern thinking about good innovation practice" (von Tunzelmann & Mbula, 2003, p. 8).

Regarding quality of research, the University Act of 1997 requires each Irish university to review the quality of their research work on a ten-year basis. This quality review system (the QAQI Program, as noted earlier) is managed by the Quality Assurance Office. The program is based on self-assessment of the unit mechanisms and peer-review by external agents, leading to a Quality Improvement Plan. The introduction of the quality standards within the United Kingdom RAE system in Britain in the 1980s had a profound effect on the subsequent development of QAQI procedures in Ireland. The importance of quality in Irish universities is clearly underlined by the funding that has been made available by the HEA under the NDP. As a collaborative measure, the governing authorities of the seven universities involved in the QAQI Program established the Irish Universities Quality Board (IUQB), which plays an important role in the selection of the agencies responsible for the periodic reviews and evaluation reports.

Benchmarking exercises have not been systematically conducted in Ireland, but some recent examples include (Tunzelmann & Mbula, 2003):

1. ICSTI carried out a benchmarking exercise on science, technology and mathematics education in 1999. Its report *Benchmarking School Science, Technology and Mathematics Education in Ireland Against International Good Practice* was released in 2000. Both
qualitative and quantitative indicators were used for this study and many weaknesses of the Irish research and education system were brought together in comparison to the benchmark selected countries (Scotland, Finland, Malaysia, and New Zealand)

2. A more recent benchmarking exercise was conducted by ICSTI whose final report called *Benchmarking Mechanisms and Strategies to Attract Researchers to Ireland* was released in 2001. The United Kingdom, United States, the Netherlands, Finland, and Denmark were the five selected benchmark countries. Recommendations for Ireland derived from this exercise were: (i) build-up centers of excellence; (ii) improve international networks and visibility of Irish universities; (iii) improve the status and remuneration of research graduates and post-doctorates; and (iv) make the move to Ireland as smooth as possible

Reporting to the Cabinet Sub-Committee, the IDC and its Joint Secretariat will oversee continuing review and evaluation, with input from the Chief Scientific Adviser and the Advisory Science Council. The well established STI evaluation capacity in Forfás will also play a valuable role in supporting the Joint Secretariat in this task. In addition, specific review mechanisms will be put in place in the context of the NDP. In reviewing the development of the strategy, there will be close liaison with stakeholder bodies including the research community and industry. A range of key indicators and targets for the strategy will be monitored, including for example: the number of new doctorates in science, engineering and technology; performance in terms of publications, and; the relative frequency with which Irish scientific publications are cited by scientific peers (von Tunzelmann & Mbula, 2003).

**Japan**

In the following, Japan’s research structure and context and evaluation model, as well as a brief historical overview of the country is presented.
Historical Overview

In 1603, a Tokugawa shogunate ushered in a long period of isolation from foreign influence in order to secure its power. For two-hundred fifty years this policy enabled Japan to enjoy stability and a flowering of its indigenous culture. Following the Treaty of Kanagawa with the United States in 1854, Japan opened its ports and began to intensively modernize and industrialize. During the late 19th and early 20th centuries, Japan became a regional power that was able to defeat the forces of both China and Russia. It occupied Korea, Formosa, and southern Sakhalin Island. In 1931 Japan occupied Manchuria, and in 1937 it launched a full-scale invasion of China. Japan attacked United States forces in 1941—triggering America’s entry into World War II—and soon occupied much of East and Southeast Asia.

After its defeat in World War II, Japan recovered to become an economic power and a staunch ally of the United States. While the emperor retains his throne as a symbol of national unity, actual power rests in networks of powerful politicians, bureaucrats, and business executives. The economy experienced a major slowdown starting in the 1990s following three decades of unprecedented growth, but Japan still remains a major economic power, both in Asia and globally. In 2005, Japan began a two-year term as a non-permanent member of the United Nations Security Council. The island chain of Japan is located in Eastern Asia, between the North Pacific Ocean and the Sea of Japan, east of the Korean Peninsula. In 2006, Japan had a population of nearly 130 million (Central Intelligence Agency, 2006).

Research Structure and Context

The administrative structure for the funding of science and technology in Japan was markedly reorganized in 2001, when the Ministry of State for Science and
Technology Policy and the Council for Science and Technology Policy (CSTP) were founded. These organizations work together to determine the national strategy for science and technology, and the policy for allocating R&D resources (they also evaluate important national R&D projects). Included in this basic scheme is the Ministry of Education, Culture, Sports, Science and Technology (MEXT), which was formed by the merger of the former Ministry of Education, Science, Sports and Culture (MESSC) and the Science and Technology Agency (STA).

The major funder of R&D in Japan is the business enterprise sector where around 75% of GERD is performed. Government funding of GERD is comparatively small, at around 18%. GERD by the private non-profit sector is around 3% (European Commission, 2006h). Funding for science and technology has increased from ¥3.58 trillion ($316 million USD) in 2005 to ¥3.8 trillion ($336 million USD) in 2006, and is distributed through the various Ministries of State in accordance with the Third Science and Technology Basic Plan.

In 2006, MEXT provided the most funding for S&T, with 63% of the total science and technology budget, the Ministry of Economy, Trade and Industry (METI) delivers 16.4% of the budget, followed by the Self Defense Agency (SDA) at 5%, the Ministry of Health, Labor and Welfare at 3.9%, the Ministry of Agriculture, Fisheries and Food at 3.7%, and the Ministry of Land, Infrastructure and Transport at 2.2% (European Commission, 2006h).

Basic research programs are delivered chiefly by the Japan Society for the Promotion of Science (JSPS) and Japan Science and Technology Agency (JST). In 2003, 9.3% of research expenditure was performed by non-academic organizations. The Independent Administrative Institutions (IAIs), which have been granted greater autonomy from government following the Independent Administrative Institution Law,
which include the Institute of Physical and Chemical Research (RIKEN), Advanced Industrial Science and Technology (AIST)\textsuperscript{279} and New Energy and Industrial Technology Development Organization (NEDO). Co-funding and public funding of private R&D in Japan is minimal (European Commission, 2006h).

In general, research funds in Japan are distributed via block funding (N. Nakamura, personal communication, March 23, 2007), however, a small proportion are distributed competitively based on research issues that have been chosen by prior evaluation (Motohashi, 2003).

**Evaluation Model**

The evaluation of research in Japan is deeply defined by the traditional position of national universities and their full support by the state. Poor quality and lack of international competitiveness in higher education and basic research were identified as the major causes of recent reforms. “National universities are still part of the Ministry of Education, but are undergoing the process of becoming more independent. They will gain the status of agents in 2004 and become more autonomous” (von Tunzelmann & Mbula, 2003, p. 28). In 2001, 56 out of 83 national research institutes were transformed into independent administrative institutions to increase the flexibility of administration, while also increasing autonomous responsibility.

The current state of R&D evaluation in Japan is that the level of R&D evaluation targets has gradually risen from individual projects toward large-scale projects and programs, and in the course of evaluating these, tools were developed, which are now being utilized in one way or another for evaluation of R&D. However, more

sophisticated means for evaluating policies (i.e., system evaluation) have not yet been established (Washington Research Evaluation Network, 2005). Despite this fact, Japan still conducts periodic research policy evaluations aimed at measuring and analyzing the effects of policies and to objectively assess them so as to provide useful information for more precise planning and implementation (Nakamura, 2006, November). Following prolonged consultation and discussion, in 2002, the Law for the Performance of Evaluating Administrative Organizations came into effect. This set out the obligations of the Ministry for Internal Affairs and Communications (SOUMU), the publication of results, and the types of evaluation that should be undertaken.

In 1997, the National Guideline on the Method for Evaluation of Government R&D was published (Prime Minister of Japan, 1997; Yamakazi, 2002). This document is related to the implementation of the Science and Technology Basic Plans, implemented on the basis of the Science and Technology Basic Law (1995). The underlying thinking regarding evaluation of R&D is premised on external evaluation to ensure clarity, third-party external evaluation, and publication of results, reflecting budgetary allocations in light of performance. Based on this guideline, it was felt that publication of results had been insufficient and lead to a stronger emphasis on this point for the Second Science and Technology Basic Plan, leading to increased emphasis and amendment through the Outline Objectives Relating to the Evaluation of National Research Activities by the Cabinet Office (European Commission, 2006h).

The CSTP prioritizes all national S&T activities annually. Use of the relatively small competitive research budget is evaluated and ranked on the basis of a four point scale (S, A, B, or C), where (European Commission, 2006h):

- S = Specially [sic] important research results
- A = Important research results
B = Efficiency requirements
C = A reconsideration of budget efficiency

The budget of science and technology policy is decided according to the value allocated. In addition, IAI activities are evaluated every mid-term period (3-5 years) by evaluation committees established by either the Cabinet Office or relevant ministries. The national universities are evaluated by the National Institute for Academic Degrees and University Evaluation (NIAD) and by special evaluation committees. Other evaluations include the evaluation of researchers by the directors of individual institutions, the self-evaluation of programs in these institutions, and sponsored evaluations such as those undertaken by the National Institute of Science and Technology Policy (NISTEP) and the Mitsubishi Research Institute (MRI) for the Science and Technology Basic Plans.

Along with the Policies for the Structural Reform of Universities in 1998, MEXT prepared the *University-based Structural Reform Plan for Revitalizing the Japanese Economy* (2000). These plans defined the future direction of research reform, with a view toward making universities more dynamic and internationally competitive. They stipulated (Ministry of Education, Culture, Sports, Science and Technology, 2000, 2002; von Tunzelmann & Mbula, 2003):

1. The realignment and consolidation of national universities should be boldly pursued
2. Management methods of the private sector should be introduced into national universities
3. A competitive mechanism with third-party evaluation should be adopted by universities.

Another major change in the Japanese research evaluation system dates from 2001, when MEXT launched the Top 30 Project. This project was designed to raise the standards of Japan’s top 30 research universities to the world’s highest levels. In each of
the above-mentioned priority research areas, funds were provided to subsidize 30 graduate-level departments. The implementation process proceeds in four steps:

1. Proposals by universities
2. Evaluations by panels of specialists
3. Selection of the top 30 departments in each area
4. Provision of funding

Under this scheme, proposals prepared by the universities are subject to a peer review by Japanese and foreign specialists who choose the top 30 departments in each priority area. The ranking scheme in this project differs from the traditional Japanese system of ranking based on the average ‘hensachi’ (deviation value of standardized test scores) of the applicants applying for university admission.

The long-term strategic goal of this program is to elevate Japanese research universities to the apex of international excellence. Nevertheless, some expert academics agree on the importance of the consideration of a separate funding framework for the program, the limitation of the government’s involvement to a support role, and the need for effective internal decision-making mechanisms within universities, in order to achieve this goal. In parallel with providing minimizing investment under the Top 30 scheme, it is seen as being highly desirable for Japan to promote the establishment of networks of competence that overcome current barriers to university-industry collaboration (European Commission, 2006h).

The primary organization responsible for quality assurance of higher education conducted research in Japan is the national government (through MEXT). Authorization and supervision by the national government is the formal base of quality assurance. This mechanism is complemented by self-evaluation. Implementation of self-monitoring/evaluation has been a required activity since the change of standard in 1999.
“Since the mid-1990s most Japanese universities have implemented the process of self-evaluation. External evaluation is being implemented via an organization called NIAD, the National Institute for Academic Degrees” (Tunzelmann & Mbula, 2003, p. 30).

NIAD was established in 1991 and was reorganized as a new body in 2000. In addition to its original degree-awarding functions it now works as a national organization for university research. This reform is the consequence of an earlier discussion raised by the University Council’s report *A Vision for Universities in the 21st Century and Reform Measures* (1998). Presently, NIAD has four major tasks (European Commission, 2006h):

1. Evaluation of education, research and other activities of universities
2. Awarding academic degrees as well as assessment and recognition/approval of programs provided by higher educational institutions
3. Conducting research on university evaluations and research on systems of academic degrees and assessment in learning adopted in other countries as well as in Japan
4. Collecting, filing, and disseminating information on university evaluation

Evaluation programs extend into three areas: university-wide thematic evaluation (UwTE); evaluation of educational activities by academic field (EEA); and evaluation of research activities by academic field (ERA). NIAD is currently evaluating “several universities as a trial...[but]...real assessment will begin soon...[and]...Japanese universities were evaluated when application to set up a new university was made, but they were not evaluated afterwards. Now, the process is about to change and Japanese universities will have to be accredited regularly (as in the United States system, not like the United Kingdom system, where the outcome of the evaluation is linked to funding)” (Tunzelmann & Mbula, 2003, p. 30).
NEDO is a semi-governmental research organization established in 1980 with an annual budget of ¥270 billion ($2.3 billion USD) as a management agency, initially for energy research. Its evaluation guidelines are set by the Office of the Prime Minister. Evaluation is carried out by external evaluators using a mix of quantitative and qualitative/narrative indicators. A key feature is the use of viva voce (i.e., spoken evidence) interviews and open panel debates. The evaluation process is divided into four stages (European Commission, 2004):

1. Ex ante assessment of the project’s potential
2. Interim evaluations during the life of the project
3. Ex post evaluation
4. A follow-up and monitoring phase which takes place five years after completion

A second, higher tier of the NEDO evaluation process is designed to help answer strategic questions relating to the receptiveness of the innovation environment, such as the fit and importance of the technology, the effectiveness of intellectual property rights, and the prospects for exploitation and commercialization of the results.

Netherlands

In the following, the Netherlands’ research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

The Dutch United Provinces declared their independence from Spain in 1579; during the 17th century, they became a leading seafaring and commercial power, with settlements and colonies around the world. After a twenty year French occupation, a
Kingdom of the Netherlands was formed in 1815. In 1830 Belgium seceded and formed a separate kingdom. The Netherlands remained neutral in World War I, but suffered invasion and occupation by Germany in World War II. A modern, industrialized nation, the Netherlands is also a large exporter of agricultural products. The country was a founding member of NATO and the EEC (now the EU), and participated in the introduction of the Euro in 1999. The Netherlands is located in Western Europe, bordering the North Sea, between Belgium and Germany, occupying 33.8 thousand square kilometers of land. In 2006, the Netherlands had a population of 16.4 million (Central Intelligence Agency, 2006).

Research Structure and Context

Presently, research in the Netherlands is funded through a dual-support system, consisting of core funding, referred to as ‘first-flow’, provided by the Dutch Ministry of Education and Science. Additional funding comes in the form of ‘second-flow’ grants come from research councils and foundations, and ‘third-flow’ contracts come from government departments and other organizations (Geuna & Martin, 2003).

In the late 1970s, growing concern about the quality and societal relevance of university research led to demands for greater accountability, and in 1979, a government White Paper—Beleidsnota Universitair Onderwijs—(Minister for Education and Science, 1979) recommended changes in the management of academic research and a system of ‘conditional funding’ was introduced in 1983 (Irvine, Martin, & Isard, 1990). Under this system funds given for teaching, ‘A-part’ funds, and research, ‘B-part’ funds, were separated and research funds were financed according to output quality. The goal of this system was to enhance quality and coherence in university research, to assess the relevance of research to society, and to have research quality translated into funding
decisions (Eiffinger, 1997).

In 1982, the Dutch Advisory Council for Science Policy (RAWB) conducted the first large-scale assessment of the quality of research at all Dutch medical facilities, partly based on bibliometric indicators (Ritger, 1983 as cited in Moed, van Leeuwen, & Visser, 1999). RAWB determined that one of the poorly performing Dutch medical facilities should be closed based on the findings from the assessment. However, the research budgets of the Netherlands institutions were more or less frozen in late 1983, and the conditional funding system lost much of its power and the facility was never closed.

In 1993, the conditional funding scheme was replaced by Hoger Onderwijs Bekostigingsmodel (HOBEK), or 'higher education funding model'. The name of this scheme was derived from the fact that the Ministry intended to fund research that had strategic relevance, meaning “relevance to society” (Geuna & Martin, 2003, p. 284). HOBEK funding allocations were weighted by teaching (23%), research (64%), and ‘inter-weavement’ (13%), based on four factors: students numbers, numbers of prior year degrees awarded, number of students completing the degree in the required four years, and numbers of graduating research students (Geuna & Martin, 2003). Unlike allocations for teaching, the HOBEK funds for research were not allocated in a normative way, and instead budgets were allocated incrementally, on a historical basis rather than according to quality (Koelman, 1998). However, in 1999 HOBEK was replaced by Stabiele Bekostiging (STABEK), or ‘stable funding’. Under the STABEK system the government approved funding for several years, which was intended to be temporary until a system which placed greater emphasis on performance could be adopted (Geuna & Martin, 2003).
Evaluation Model

Prior to the implementation of STABEK, in 1992, the Netherlands' Minister of Education and thirteen universities agreed that the Association of the Netherlands Universities (VSNU) should develop an external evaluation system to compliment preexisting, internal quality controls (van Steen & Eijffinger, 1998). Although the Netherlands a long history of evaluating its publicly-funded research, which has been referred to as a 'patchwork system' (Rip & van der Meulen, 1995) in earlier incarnations, beginning in 1993, the Netherlands universities began to have their research programs systematically evaluated by international committees of independent experts (van Steen & Eijffinger, 1998). These evaluations were conducted across the existing twenty-seven disciplines and programs, each by a different committee, whose members were predominately foreign to insure impartiality, although the committee chairs were often Dutch. Furthermore, unlike the United Kingdom’s RAE, the Dutch phased their evaluations periods of four to six years, as opposed to simultaneous evaluation of all disciplines as is the practice in the United Kingdom (RAE 2008, 2005).

A pilot evaluation was conducted in 1993 for a select group of disciplines by committees of five to seven experts, selected by VSNU in cooperation with the Royal Netherlands Academy of Arts and Sciences (KNAW). These committees evaluated the pilot disciplines over periods of five years on the basis of (Geuna & Martin, 2003):

1. Academic staff
2. Program mission and research plan
3. Content of programs and main results
4. Publications
5. Five selected publications
6. Other indicators of quality and reputation such as patents and invited lectures

These committees conducted extensive site visits and interviews of program leaders and directors, and VSNU also commissioned bibliometric analyses to enhance the committee peer review process (VSNU, 1996). Each of the disciplines and programs was assessed by the committees on the following four dimensions (Geuna & Martin, 2003):

1. **Scientific quality.** Originality of ideas and methods, importance of the discipline, impact, and prominence

2. **Scientific productivity.** Inputs and outputs (staff and funds are inputs; while outputs are number and nature of publications, dissertations, patents, and invited lectures

3. **Scientific relevance.** Relevance to the advancement of knowledge and technology and social consequences

4. **Long-term viability.** For research, publication, coherence, and continuity of research

In turn, these were then converted to one of the five ratings presented in Table 5

(VSNU, 2003, p. 25)
Table 5

Netherlands’ Standard Evaluation Protocol
for Public Research Organizations Five-Point Scale

<table>
<thead>
<tr>
<th>Scale</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>Work that is at the forefront internationally, and which most likely will have an important and substantial impact in the field. Institute is considered an international leader.</td>
</tr>
<tr>
<td>Very good</td>
<td>Work that is internationally competitive and is expected to make a significant contribution; nationally speaking at the forefront in the field. Institute is considered international player, national leader.</td>
</tr>
<tr>
<td>Good</td>
<td>Work that is competitive at the national level and will probably make a valuable contribution in the international field. Institute is considered internationally visible and a national player.</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>Work that is solid but not exciting, will add to our understanding and is in principle worthy of support. It is considered of less priority than work in the above categories. Institute is nationally visible.</td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>Work that is neither solid nor exciting, flawed in the scientific and technical approach, repetitions of other work, etc. Work not worth pursuing.</td>
</tr>
</tbody>
</table>

In 2003, the VSNU, KNAW, and the Netherlands Organisation for Scientific Research undertook efforts to introduce a new system of quality control for university and para-university research, based on the report *Kwaliteit Verplicht* (the obligation of quality) of the working group Kwaliteitszorg Wetenschappelijk Onderzoek (2001). A characteristic feature of this new system was that each research school or research institute would be assessed in terms of its academic standards every six years by an international group of independent experts and required to conduct a self-evaluation every three years. Therefore, the system was not only to be used to allocate funds, but also to develop strategies. The revised system was intended to serve three purposes (VSNU, 2003):

1. *Improvement* of the quality of research according to international standards of quality and relevance

2. *Improvement* of research management and leadership
In the same year the VSNU assessments for each discipline were discontinued, and the system placed its emphasis on institutions, rather than disciplines or programs. In part, this change can be attributed to ongoing efforts to achieve a division and concentration of tasks via the formation of “research schools, that is, collections of research groups, mostly from different universities and active in the same specialty” (Moed, van Leeuwen, & Visser, 1999, p. 61). The revised Standard Evaluation Protocol 2003-2009 (VSNU, 2003) was completed in 2003 and the evaluations are to be both retrospective and prospective, as reflected in the assessment criteria for past performance and future plans (Royal Netherlands Academy of Arts and Sciences, 2002).

This system aims at operating with the least possible burden on researchers (a self-evaluation once every three years and external evaluation every six years) and it also encourages accessibility through the Internet of certain relevant data by the generation of a national research information system. With the new procedure the Netherlands is giving a more central role to self-evaluation. The evaluation process is composed of nine steps, as follows (von Tunzelmann & Mbula, 2003):

1. Planning and timetable for all research institutes, including a protocol draft for each specific evaluation
2. Protocol for the specific external evaluation
3. Selection of the evaluation committee
4. Self-evaluation
5. Evaluation committee’s working program
6. Evaluation Report, including a review of the entire institute and a review of each research program
7. Conclusions by the Board, based on the self-evaluation document and the final evaluation report
New Zealand

In the following, New Zealand's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

The Polynesian Maori reached New Zealand in about 800 AD. In 1840, their chieftains entered into a compact with Britain, the Treaty of Waitangi, in which they ceded sovereignty to Queen Victoria while retaining territorial rights. A series of land wars between 1843 and 1872 ended with the defeat of the native peoples. The British colony of New Zealand became an independent dominion in 1907 and supported the United Kingdom militarily in both World Wars. In recent years, the government has sought to address longstanding Maori grievances. The islands of New Zealand had a population of more than 4.0 million in 2006 (Central Intelligence Agency, 2006).

Research Structure and Context

For many years, research in the tertiary education sector of New Zealand was funded mainly through public tuition subsidies based on the number of equivalent-fulltime students (EFTS) and with weightings for different courses based, at least to some degree, on the cost of provision. Tertiary education organizations (TEO) secured research funds from the Foundation for Research, Science and Technology (FRST), the Health Research Council (HRC), the Marsden Fund (managed by the Royal Society of New Zealand), government departments and agencies, and the private sector.
Nevertheless, most TEOs were heavily dependent upon EFTS funding in order to support their research activities. This meant that certain research programs were vulnerable to large shifts in student demand. It has also meant that the volume of research in particular subject areas was determined more by the pattern of student demand than by the quality of research being undertaken. In the late 1990s, a portion of the EFTS subsidies for degree-level programs was designated for research in the form of degree ‘top ups’ and the subsidy rates for different course categories were adjusted. This did not, however, alter the fundamental nature of the research funding system in the tertiary education sector; nor did it address the underlying weaknesses.

Since 1999, significant efforts have been made to improve the tertiary funding regime in the interests of encouraging and rewarding excellence. The first major step in this process was the government’s decision in 2001 to fund the creation of a number of centers of research excellence (CORE) within the tertiary sector. Initially five COREs were established, with funding commencing in 2002. An additional two were funded in 2003. They are being reviewed and possibly expanded in 2007.

A second key step was the establishment of the Performance-Based Research Fund (PBRF). This new program allocated research funding to participating TEOs for the first time in 2004, and with the funding allocation comes the periodic assessment of research quality together with the use of two performance indicators. Between 2004 and 2007 all the funding that is currently distributed via the ‘top ups’ will gradually be transferred to the PBRF. Additionally, the government has allocated significant new funding which will be phased-in over the next three years, so that in 2007 close to $33 million New Zealand dollars (NZD; $22 million USD) of extra funding will be available to participating TEOs. According to current forecasts, it is estimated that in 2007 approximately $185 million NZD ($125 million USD) will be allocated through the
PBRF. This will make the PBRF the largest single source of research funding for New Zealand’s tertiary education sector.

The government’s decision in mid 2002 to introduce the PBRF marked the culmination of many years of vigorous debate over the best way of funding research in the country’s tertiary education sector. In 1997, the previous national-led government proposed a new system for research funding and subsequently appointed a team of experts to consider the options. For various reasons, little progress was made. In 2001, the Tertiary Education Advisory Commission (TEAC, since changed to TEC), which was appointed by the Labour-Alliance government, recommended the introduction of the PBRF as a central component of a new funding regime for the tertiary sector.

The TEAC proposal was the product of detailed consultation with the tertiary education sector and comparative analysis of various overseas approaches to the funding of research. In essence, TEAC recommended a mixed model for assessing and funding research, which on the one hand, incorporated an element of peer review, and on the other, several quantifiable performance indicators (Tertiary Education Commission, 2003a, 2004a, 2004b, 2004c).

In response to the TEAC report, the government established a working group of sector experts in mid 2002 to develop the detailed design of a new research assessment and funding model for the tertiary sector. The report of the Working Group on the PBRF, Investing in Excellence, was published in December 2002 and approved by the Cabinet.

The Working Group endorsed the key elements of the funding model proposed by TEAC, including the periodic assessment of research quality by expert panels and the use of two performance indicators. It also supported TEAC’s idea of using individuals as the unit of assessment, rather than academic units. It did, however, recommend that the
funding formula have different weightings from those proposed by TEAC, and it developed a comprehensive framework for assessing the research performance of individual staff.

**Evaluation Model**

The focus of the PBRF is on revealing and rewarding *researcher excellence and excellent research*, as defined in terms of “producing and creating leading-edge knowledge, applying that knowledge, disseminating that knowledge to students and the wider community, and supporting current and potential colleagues to create, apply and disseminate knowledge” (von Tunzelmann & Mbula, 2003, p. 36).

The PBRF funding formula is based on the following three elements (Tertiary Education Commission, 2005):

1. **Quality evaluation (QE).** The assessment of the research quality of TEO staff members, based on peer review
2. **Postgraduate research degree completions (RDC).** The number of postgraduate research-based degrees completed in the TEO
3. **External research income (ERI).** The amount of income for research purposes received by the TEO from external sources

The weightings in the funding formula for the three elements are: \( \text{QE} = 60\% \), \( \text{RDC} = 25\% \), and \( \text{ERI} = 15\% \). Funding is distributed to each participating tertiary education provider as a block grant from a fixed PBRF funding pool proportional to their performance in the three elements. The PBRF is, then, an extremely competitive system in which increases in funding to one provider necessarily lead to reductions to the others.

All degree-granting tertiary education providers and all academics within them who undertake research and/or degree-level teaching are eligible, but not required, to
participate in the QE process. If they do not participate they are automatically assigned a quality category of R (meaning no research dollars). The evaluation is undertaken by peer review panels, using information provided by each eligible researcher in an individual Evidence Portfolio (EP) submitted to the TEC by their employing institution. The QE takes into account a range of factors, including research outputs, esteem factors, and contributions to the development of both new researchers and a vital high-quality research environment.

For New Zealand’s QE research is defined as:

...original investigation undertaken in order to contribute to knowledge and understanding and, in the case of some disciplines, cultural innovation or aesthetic refinement. It typically involves enquiry of an experimental or critical nature driven by hypotheses or intellectual positions capable of rigorous assessment by experts in a given discipline. It is an independent, creative, cumulative and often long-term activity conducted by people with specialist knowledge about the theories, methods and information concerning their field of enquiry. Its findings must be open to scrutiny and formal evaluation by others in the field, and this may be achieved through publication or public presentation. In some disciplines, the investigation and its results may be embodied in the form of artistic works, designs or performances. Research includes contribution to the intellectual infrastructure of subjects and disciplines (e.g., dictionaries and scholarly editions). It also includes the experimental development of design or construction solutions, as well as investigation that leads to new or substantially improved materials, devices, products or processes (Tertiary Education Commission, 2005, p. 20).

Ten principles guide New Zealand’s PBRF, and the QE, which are (Tertiary Education Commission, 2005):

1. **Comprehensiveness.** The PBRF should appropriately measure the quality of the full range of original investigative activity that occurs within the sector, regardless of its type, form, or place of output.

2. **Respect for academic traditions.** The PBRF should operate in a manner that is consistent with academic freedom and institutional autonomy.
3. **Consistency.** Evaluations of quality made through the PBRF should be consistent across the different subject areas and in the calibration of quality ratings against international standards of excellence.

4. **Continuity.** Changes to the PBRF process should only be made where they can bring demonstrable improvements that outweigh the cost of implementing them.

5. **Differentiation.** The PBRF should allow stakeholders and the government to differentiate between providers and their units on the basis of their relative quality.

6. **Credibility.** The methodology, format and processes employed in the PBRF must be credible to those being assessed.

7. **Efficiency.** Administrative and compliance costs should be kept to the minimum consistent with a robust and credible process.

8. **Transparency.** Decisions and decision-making processes must be explained openly, except where there is a need to preserve confidentiality and privacy.

9. **Complementarity.** The PBRF should be integrated with new and existing policies, such as charters and profiles, and quality-assurance systems for degrees and degree providers.

10. **Cultural inclusiveness.** The PBRF should reflect the bicultural nature of New Zealand and the special role and status of the Treaty of Waitangi, and should appropriately reflect and include the full diversity of New Zealand’s population.

The EPs are the key component of the PBRF and form the basis of the QE measure. Each submitted EP must consist of three elements (Tertiary Education Commission, 2005).

1. **Research outputs** (RO). The outputs of a staff member’s research; separated into four nominated ROs (which must be available for scrutiny by the panels) and up to 30 others referenced

2. **Peer esteem** (PE). An indication of the quality of the research of the staff member, as recognized by their peers

---

280 Evidence portfolio (EP) requirements differ slightly for Māori, Pasifika (Pacific), and new or emerging researchers in the 2006 round.
3. **Contribution to the research environment** (CRE). The staff member’s contribution to a vital high-quality research environment, both within the TEO and beyond it.

EPs are submitted to one of twelve disciplinary panels (e.g., biological sciences, humanities and law, mathematical and information sciences and technology), which consist of Panel Members, a Panel Chair, and a Panel Secretariat. The panels review each EP in detail and then assign preparatory and preliminary scores for each of the three components of the EP, followed by the full panel review of those scores, and finally assign a Quality Category to each EP via a process of holistic assessment.

The scoring system used in the QE is based on allocating four Quality Categories (A [7-6], B [5-4], C [3-2], and R [1-0]), which are differentially weighted for funding purposes and points for each of the three components of the EP on a scale from 0 to 7, where 0 = no evidence of a research platform on that component. Each of the three components is then weighted from 70% to 15% as presented in Table 6.

<table>
<thead>
<tr>
<th>Component</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research output (RO)</td>
<td>70%</td>
</tr>
<tr>
<td>Peer esteem (PE)</td>
<td>15%</td>
</tr>
<tr>
<td>Contribution to the research environment (CRE)</td>
<td>15%</td>
</tr>
</tbody>
</table>

Table 7 illustrates an example of how the total weighted score is calculated for the RO part of the model, in the case of EP scores where RO = 4, PE = 6, and CRE = 5.
Table 7
New Zealand's Quality Evaluation Total Weighted Score Calculation

<table>
<thead>
<tr>
<th>EP component</th>
<th>Raw score (0-7)</th>
<th>Weighting</th>
<th>Weighted score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RO</td>
<td>4</td>
<td>70%</td>
<td>280</td>
</tr>
<tr>
<td>PE</td>
<td>6</td>
<td>15%</td>
<td>90</td>
</tr>
<tr>
<td>CRE</td>
<td>5</td>
<td>15%</td>
<td>75</td>
</tr>
<tr>
<td><strong>Total weighted score</strong></td>
<td></td>
<td></td>
<td><strong>445</strong></td>
</tr>
</tbody>
</table>

Total weighted scores are then used to assign a Quality Category, as presented in Table 8. As shown in the table, these categories are differentially stepped (i.e., increasingly difficult to achieve), where an A has a 100-point range and the B, C, and R categories have 200-point ranges.

Table 8
New Zealand's Quality Evaluation Total Weighted Scores and Quality Categories

<table>
<thead>
<tr>
<th>Total weighted score</th>
<th>Quality category</th>
</tr>
</thead>
<tbody>
<tr>
<td>600-700</td>
<td>A</td>
</tr>
<tr>
<td>400-599</td>
<td>B</td>
</tr>
<tr>
<td>200-399</td>
<td>C and C(NE)</td>
</tr>
<tr>
<td>&lt;200</td>
<td>R and R(NE)</td>
</tr>
</tbody>
</table>

In the example above, where the total weighted score was 445, the case would be assigned to the B quality category, which is described as: “The staff member has produced research outputs of a high quality, acquired recognition by peers for their research at least at a national level, and made a contribution to the research environment beyond their institution and/or a significant contribution within their institution,” as shown in Table 9.
Table 9

New Zealand’s Quality Evaluation Quality Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The staff member has produced research outputs of a world-class standard, established a high level of peer recognition and esteem within the relevant subject area of their research, and made a significant contribution to the New Zealand and/or international research environments.</td>
</tr>
<tr>
<td>B</td>
<td>The staff member has produced research outputs of a high quality, acquired recognition by peers for their research at least at a national level, and made a contribution to the research environment beyond their institution and/or a significant contribution within their institution.</td>
</tr>
<tr>
<td>C</td>
<td>The staff member has produced a reasonable quantity of quality-assured research outputs, acquired some peer recognition, and made a contribution to the research environment within their institution.</td>
</tr>
<tr>
<td>C(NE)</td>
<td>A new or emerging researcher would be expected to have produced a reasonable platform of research, as evidenced by having: either a) completed their doctorate of equivalent qualification and produced at least two quality-assured research outputs or b) produced research outputs equivalent to a doctorate and at least two quality-assured research outputs.</td>
</tr>
<tr>
<td>R</td>
<td>Does not demonstrate the quality standard required for a C Quality Category or higher.</td>
</tr>
<tr>
<td>R(NE)</td>
<td>Does not demonstrate the quality standard required for a C(NE) Quality Category of higher.</td>
</tr>
</tbody>
</table>

The evaluation process used by the twelve disciplinary panels involved an eight step procedure, starting with the preparatory scoring and concluding with the assignment of final Quality Categories, as shown below (J. Hattie, personal communication, February 18, 2006):

1. **Preparatory** scores for each of the three components by two members of each panel
2. **Cross-referential** scores for each of the components
3. **Preliminary** scores for each of the three components by the panel in light of 1 and 2
4. **Indicative** Quality Categories based on the preceding sets of scores
5. **Calibrated panel** scores for each of the three components based on
the calibration of the preceding set of scores

6. *Calibrated Panel* Quality Categories based on the calibrated scores

7. *Holistic* Quality Categories based on a holistic judgment of each EP

8. *Final* Quality Categories

The first QE was conducted in 2003, and was the first large-scale, systematic assessment of the quality of research in New Zealand (Tertiary Education Commission, 2004b). Of the 45 PBRF-eligible TEOs, twenty-two participated in the 2003 QE, which were comprised of eight universities, two polytechnics, four colleges of education, one wānanga, and seven private training establishments. Under the agreed procedures for this round, participating TEOs undertook an initial assessment of the Evidence Portfolios (EP) prepared by their PBRF-eligible staff and assigned each portfolio one of four possible Quality Categories (A, B, C, and R). Those assigned an A, B, or C were submitted to the TEC for assessment by a peer review panel. Data were supplied to the TEC on the EPs that were assigned an R.

Of the 8,013 PBRF-eligible staff in the participating TEOs, 5,771 had their EPs assessed by a peer review panel, covering 41 designated subject areas. The work of these expert panels was overseen by a Moderation Panel comprising the twelve Panel Chairs and an independent chair. Altogether, there were 165 Panel Chairs and Panel Members, of whom 33 were from overseas.

The experience gained in the 2003 QE was used to provide input into the redesign of the PBRF in preparation for the 2006 QE. Following consultation with the sector, a Sector Reference Group (SRG) was formed to consider the issues highlighted by the implementation of the 2003 QE, the *Phase 1 Evaluation of the PBRF* (Web Research, 2004), and the reports of the peer review panels (e.g., Alcorn, et al., 2006).
The total administrative and compliance costs of implementing the PBRF in 2003 and conducting the 2003 QE was estimated between $14 and $21 million NZD (between $10 and $15 million USD) or between 14% and 21% of the total PBRF funding for the period 2004-2006 allocated by the 2003 QE, compared to less than 1% for both the United Kingdom and Hong Kong exercises. Although, these costs were anticipated:

...this is, on any measure, a high cost. However, the combination of chosen design elements, complexity, newness and demands on staff time ensured that the costs of the implementation of the PBRF and the simultaneous conduct of the 2003 Quality Evaluation were always going to be high. They were also high as a proportion of the PBRF funding allocated because the PBRF funding allocated by the 2003 Quality Evaluation and the RDC and ERI results, was the smallest it may ever be (Web Research, 2004, p. 78).

Moreover, the projected costs, including administration and compliance, in 2006 will be substantially lower, and estimated at between only 1% and 2.3% for the period 2007 to 2012 (Web Research, 2004). For the second round (2006), major changes included new categories for 'new and emerging' researchers, a moderation panel of three to oversee the process across all panels, increases in the numbers of EPs that were cited, and improvements in the moderation process to provide more information to the panels prior to their meeting. The PBRF has also allocated substantial funds for metaevaluation of the QE, but thus far the quality of the metaevaluation has been less than desired (J. Hattie, personal communication, March 15, 2007).

Poland

In the following, Poland's research structure and context and evaluation model, as well as a brief historical overview of the country is presented.
Historical Overview

Poland is an ancient nation that was conceived near the middle of the 10\textsuperscript{th} century. Its golden age occurred in the 16\textsuperscript{th} century. During the following century, the strengthening of the gentry and internal disorders weakened the nation. In a series of agreements between 1772 and 1795, Russia, Prussia, and Austria partitioned Poland amongst themselves. Poland regained its independence in 1918 only to be overrun by Germany and the Soviet Union in World War II. It became a Soviet satellite state following the war, but its government was comparatively tolerant and progressive. Labor turmoil in 1980 led to the formation of the independent trade union 'Solidarity' that over time became a political force and by 1990 had swept parliamentary elections and the presidency. A 'shock therapy' program during the early 1990s enabled the country to transform its economy into one of the most robust in Central Europe, but Poland still faces the lingering challenges of high unemployment, underdeveloped and dilapidated infrastructure, and a poor rural underclass.

Solidarity suffered a major defeat in the 2001 parliamentary elections when it failed to elect a single deputy to the lower house of Parliament, and the new leaders of the Solidarity Trade Union subsequently pledged to reduce the Trade Union's political role. Poland joined NATO in 1999 and the EU in 2004. With its transformation to a democratic, market-oriented country largely completed, Poland is an increasingly active member of Euro-Atlantic organizations. In 2006, Poland had a population of 38.5 million (Central Intelligence Agency, 2006).

Research Structure and Context

In 1991, Poland initiated a new system for managing research, led by the Committee for Scientific Research (CSR). The Chairman of the CSR is appointed by
Parliament, and two-thirds of its members are researchers elected by the scientific community, with the remainder being Ministers. The CSR is responsible for Polish science policy and for the distribution of research funds through competitive channels (European Commission, 2006j; Geuna & Martin, 2003).

In 2006, the science budget was estimated at 3.3 billion Polish Zlotych (PLN; $1.1 billion USD), and according to the implementing document of the Polish National Reform Program (NRP) 2005-2008, the budget should increase to 4.9 billion PLN ($1.6 billion USD) by 2008. The main part of the budget is distributed on an institutional basis (statutory funding for research activities and infrastructure), including (European Commission, 2006j):

1. Commissioned projects
2. Development research projects
3. Goal oriented research projects
4. Special research projects
5. Long-term governmental programs
6. Technology credits
7. Status of R&D centers
8. Fiscal incentives
9. Strengthening cooperation between R&D sphere and economy

On the basis of the National Framework Program (KPR), the minister of science and higher education launches calls for commissioned projects. One of the strengths of KPR is that it established 38 fields of research in nine strategic research areas. The development research projects finance applied research and development activities aimed at practical application, whereas the goal oriented research projects concern all areas of applied research, development activities, industrial research, and pre-competitive
research indispensable to the implementation of project. The final outcome of such projects should be the implementation of product or technology as well as economic or social application. The special research projects include scientific research or development activities, which are part of international program but cannot be financed from international financial resources. The long-term government programs finance research projects in specific branch of industry (European Commission, 2006j).

The process of transformation of the Research and Development Units (JBR) was triggered by the decrease of public funding and low demand for R&D from industry that began to be privatized. As a result, the JBR had to compete on small-scale projects, but the majority of them have adjusted to the new economic climate by reducing their size and using their assets to generate revenues to ensure survival. Currently, there are 197 JBR, employing 22,991 persons. Poland’s research system has been evolving rapidly in the recent years; however, its large research base continues to be mainly financed through the national budget. According to the recent statistics, 61.7% of the R&D activities incurred by public research organizations were financed by the national budget (European Commission, 2006j).

There have also been considerable changes introduced into the governance system of Polish R&D recently. After the democratic transformation, starting from the beginning of 1990s, the R&D system was subsequently transformed. The key step was the establishment of the State Committee for Scientific Research (KBN) in the 1990s that was a body elected by all scientific community with the mandate to carry out R&D policy for Poland, including all financial decisions. The institutional arrangements have not lived up to their expectations, and in 2004 another considerable changed was introduced to the R&D system by establishing the Ministry of Science and Information Technologies (MNiI), which was transformed into the Ministry of Science and
Education (MNiE), and more recently into the Ministry of Science and Higher Education (MNiSW).

Evaluation Model

Like Hungary, and throughout Eastern Europe after 1945, national research systems in Poland were based on the Soviet tripartite model in which universities focused on teaching, basic research was conducted in Academy of Sciences’ institutes, and applied research was carried out in institutes under the various ministries (Frankel & Cave, 1997). Governments also adopted the Soviet system for funding research largely through block grants to institutes, a system in which the scientist-administrator had great power over the internal distribution of funds. Favoritism and political connections often gave rise to poor quality researchers and research teams with the right connections being funded, while high-quality researchers and research teams were under-funded (Geuna & Martin, 2003).

As mentioned previously, legislation in 1991 set up a new system for managing research, led by the CSR. All institutions, including university faculties, compete for funds through the CSR. Polish universities compete for funds on the basis of student numbers and through two CSR schemes. The first is a grant system for individuals and teams, based on open competition and peer-review. The second is so-called ‘statutory funding,’ which is distributed to faculties within universities on the basis of ex post evaluations. Each year, institutions submit their past year’s achievements and a research plan for the coming year. Assessments are conducted by expert panels who assign institutions to a category. Allocations are decided by the CSR’s Committee for Basic and Applied Research.
Until recently, funding levels were determined by an algorithm using a combination of quantitative and qualitative factors. However, the latter were criticized for their subjectivity, and in 1998 a new formula was introduced. This new 'parametric system' is almost entirely quantitative. It consists of a number of points given for performance, $R_p$, and for general results, $R_g$. The total number, $R = R_p + R_g$, is divided by the number of staff ($N$) to yield an indicator of effectiveness ($E$). This is the basis for, every three years, classifying institutions into one of five grades, and for determining their level of funding. Recently, however, the new formula has also been challenged, and it too will probably be changed (Geuna & Martin, 2003).

Research funds administered by CSR are an important source for universities. In addition to general university funds allocated on the basis of education profile (e.g., student numbers), universities also compete for funds with other institutions through two CSR schemes. The first is a grant system for individuals and research teams based on open competition; applications are peer-reviewed by panels of active researchers. The other is the so-called statutory funding that is distributed to faculties within universities (e.g., mathematics, sociology) on the basis of an ex post evaluation. Each year, institutions submit applications that include a record of the past year's achievements and a research plan for the coming years. The assessment is conducted by expert panels of scientists who assign institutions to a category ($A =$ best and $C =$ poorest). The final funding allocation is decided by the CSR (Geuna & Martin, 2003).

Like other new EU members, evaluation in Poland is relatively a new practice. Officially, the Science Council may carry out evaluations, but no evaluations have been performed so far. On the other hand, programs, which have been co-funded by the EU SFs are beginning to be evaluated. Recently, three evaluations have been completed regarding the implementation of Operational Program Increasing Competitiveness of
Economy 2004-2006 in order to give an appraisal of implementation progress, identify organizational barriers from the perspective of beneficiaries and from the perspective of implementing authorities. The evaluations have been conducted by external independent experts and are publicly available (European Commission, 2006j).

However, these assessments are not "fully-fledged evaluations including analysis of research results and impacts, thus they should be considered rather as pilot studies. This can be partially explained by the fact that the majority of projects are still ongoing and the outcomes are unknown. Taking into account that there is a requirement to conduct evaluations of EU SF programs, one can expect the increase of requests for evaluation studies" (European Commission, 2006j).

Sweden

In the following, Sweden’s research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

A military power during the 17th century, Sweden has not participated in any war in almost two centuries and an armed neutrality was preserved in both World Wars.

281 Like the Czech Republic, Poland has been particularly influenced by the EU’s FP7, which states that “knowledge lies at the heart of the EU’s Lisbon Strategy to become the ‘most dynamic competitive knowledge-based economy in the world. The ‘knowledge triangle’—research, education and innovation—is a core factor in European efforts to meet the ambitious Lisbon goals. Numerous programmes, initiatives, and support measures are carried out at EU level in support of knowledge. The FP7 bundles all research-related EU initiatives together under a common roof playing a crucial role in reaching the goals of growth, competitiveness and employment; along with a new Competitiveness and Innovation Framework Programme (CIP), Education and Training programmes, and Structural and Cohesion Funds for regional convergence and competitiveness. It is also a key pillar for the ERA. The broad objectives of FP7 have been grouped into four categories: Cooperation, Ideas, People and Capacities. For each type of objective, there is a specific programme corresponding to the main areas of EU research policy. All specific programmes work together to promote and encourage the creation of European poles of (scientific) excellence” (CORDIS, 2007b).
Sweden's long-successful economic formula of a capitalist system interlarded with substantial welfare elements was challenged in the 1990s by high unemployment and in 2000-2002 by the global economic downturn, but fiscal discipline over the past several years has allowed the country to weather economic vagaries. Indecision over the country's role in the political and economic integration of Europe delayed Sweden's entry into the EU until 1995, and waived the introduction of the Euro in 1999. Sweden had a population of 9.0 million in 2006 (Central Intelligence Agency, 2006).

Research Structure and Context

The Swedish government annually invests some 25 billion Swedish Krona (SEK; $3.5 billion USD) in R&D and six semi-public research foundations contribute another SEK1.5 billion ($226 million USD). Of the Swedish government's direct R&D investment, 56% goes to curiosity-driven research and 42% to mission-oriented R&D (20% to defense-related research and 22% to non-defense-related research). The clear majority of the investment in curiosity-driven research (SEK11 billion; $1.5 billion USD) is transferred directly to the Swedish universities and the university colleges and the remainder is funneled through three research councils. A range of sector agencies manages the investment that is not directly defense-related. The main flows of public R&D funding from the government are through the Ministry of Education, Research and Culture (52% of government R&D funding), the Ministry of Defense (20%), and the Ministry of Industry, Employment and Communications (13%). The main beneficiaries of government R&D funding are universities and university colleges, which ultimately receive over 60% of the total, and industry, which receives around 20%. In recent years, the defense sector has been restructured, resulting in a dramatic reduction in government R&D funding, meaning that the relative proportion of R&D funds to this sector is likely
In Sweden, universities and university colleges ultimately receive over 60% of the government's investments in R&D, of which 43% is directly disbursed from the government. The remainder of the government's investments in curiosity-driven R&D is managed by three research councils (European Commission, 2005, 2006k):

1. Swedish Research Council (VR), supporting basic research in all fields of science
2. Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS)
3. Swedish Council for Working Life and Social Research (FAS)

VR clearly dominates with a budget approximately double that of FORMAS and FAS combined. With few exceptions, grants from the research councils go to universities. Curiosity-driven R&D is also funded by six semi-public research foundations. The two dominant research foundations are the Swedish Foundation for Strategic Research (SSF; supporting research in natural science, engineering and medicine) and the Knowledge Foundation (KKS; supporting research at new universities and university colleges) and to a limited extent by sector agencies, such as the Swedish Governmental Agency for Innovation Systems (VINNOVA; supporting research and development in technology, transport, and working life).

All basic research funding from research councils and sector agencies, as well as most from semi-public research foundations, is allocated through peer-review systems; however, these funds are usually distributed as “a lump sum to each university which then has discretion as to its internal distribution” (Geuena & Martin, 2003, p. 14). With few exceptions, funding goes to Swedish organizations. The most recent research policy bill defines three prioritized research areas, namely life science, engineering and sustainable development to which additional funds will be allocated to further reinforce
previous bills' emphasis on support of these areas (European Commission, 2006k):

1. Additional funds allocated to the area of life science (medical research) will be distributed through VR and to a lesser degree FAS, which both use competitive calls and peer reviews.

2. Additional funds allocated to the area of engineering (technological research) will be distributed through VR, VINNOVA and to a lesser degree the Swedish National Space Board (SNSB), which all use competitive calls and peer reviews.

3. Additional funds allocated to the area of sustainable development will be distributed through FORMAS, VR and to a lesser degree VINNOVA, which all use competitive calls and peer reviews.

In addition to these prioritized research areas, the bill provides additional long-term funding for centers of excellence in both curiosity-driven and mission-oriented research, which is allocated through several research councils and sector agencies. Evaluation processes feature competitive calls using international peers. Following ramp-up, this funding will total SEK300 million ($42 million USD) per annum from 2008 and onwards and grants will be for up to SEK10 million ($1.4 million USD) per annum, per center, for a period of up to 10 years (European Commission, 2006k).

The bill further earmarks funds to facilitate academic careers for young people in order to prepare for the fact that 45% of all teaching and research staff at Swedish universities will retire within 15 years. The bulk of these funds are allocated directly to universities to facilitate postgraduate education and postdoctoral positions. Further funding is targeted for competitive calls for postdoctoral positions and for a new type of graduate school and is distributed through the three research councils and VINNOVA. The nature of such funding and the ways in which it is disbursed varies significantly depending on the institute's field of responsibility. In some cases, funds are directly disbursed from the three ministries, in others through sector agencies and semi-public research foundations. A common trait is that base funding is provided on a level
considerably lower than in all comparable nations, meaning that Swedish institutes compete internationally on very uneven terms (European Commission, 2006k).

**Evaluation Model**

In Sweden, initiatives to perform evaluations of research policies and programs are usually taken by funding agencies. Thus, the research councils, VINNOVA, and the semi-public research foundations initiate evaluations of their own programs on a regular basis. It is common that larger programs include ex ante, half-time, and ex post evaluations. Evaluations of research programs often include peer reviews to assess scientific quality (European Commission, 2006k; Geuna & Martin, 2003).

Performance-based research funding has not been implemented in Sweden and presently, a new funding system is being designed which is based on educational tasks negotiated between the Ministry and the individual institutions in which the three-year objectives of the HEIs are generally stated. These contracts usually contain the following elements (von Tunzelmann & Mbula, 2003):

1. The number of credit points that the institutions are required to award
2. The total number of FTE students
3. The fields of study in which the number of students is to increase or decrease
4. The programs in which the share of women or men is to increase
5. The follow-up to be made in the annual report
6. Special assignments

In spite of this wider regarding of quality issues in the design of new programs, the important elements in the Swedish contracts remain the regulation of student numbers and number of candidates and credit points awarded. The Swedish Natural
Science Research Council (NFR) is a governmental body which is the main funding agency for support to basic research in the natural sciences in Sweden. The support is varied in terms of providing grants for research, initiating research projects, establishing research posts, granting scholarships, promoting international cooperation, supporting scientific publications and informing the general public about research, for example. The approach is essentially what is known as 'bottom up' (i.e., not steered, nor restricted in any way), though some money may be reserved for priority areas selected by the NFR or chosen on the recommendation of the Ministry of Education and which naturally reflect current government policy. The subjects of the priority areas are usually of an interdisciplinary nature. The total amount of money involved means that both the government and the taxpayer, from whom the money comes, hold NFR accountable as to how it is spent. The financial accountability is carried out in the normal fashion by audit, but the scientific accountability is done through international reviews and evaluations and approximately SKR1 million ($142 thousand USD) is set aside each year in the budget for this purpose (von Tunzelmann & Mbula, 2003).

The word international is taken to mean that experts outside of the Swedish academic/scientific system are invited to participate. The invited experts are usually not only renowned in their own research field, but have also played a leading role in international activities whether it be in other research funding organizations, councils of scientific societies, as an international scientific advisor, a consultant to the UN, an international journal editor, or in international correlation programs or international expeditions, for example (Guy-Ohlson, 1997). Reviews of whole scientific areas (i.e., disciplines), for example, chemistry (NFR, 1995a), biology (NFR, 1995b), mathematics (NFR, 1995c), and earth sciences (NFR, 1995d), have been carried out and published.
The aim of these reviews was not only to look at the current role and status of the subjects in Sweden, even in a social context, but also to pin-point problems and make, where possible, recommendations for changes to the government, to the NFR Council, and to the universities and academic system as a whole. The reviews were also to determine if there were justifiable reasons why certain aspects of research were missing in Sweden from a scientific discipline. When reviews of this kind are made they are generally on the basis of reference groups with no individual persons, teams nor projects being named or mentioned but only the subject in its entirety being examined (Guy-Ohlson, 1997). By contrast, the evaluation of subjects usually includes evaluation of research projects and the individual scientists or teams who have received funding to carry them out.

These subject evaluations of disciplines are lead by a steering group, called the Programme Committee (PC). In earth sciences, for example, there are at present 18 members of the PC and each of them usually serves for a period of three years which is consecutively renewable only once, for a further three years. The members represent scientific expertise in their own sub-field of the discipline and have proven track records. Among other things it is for them to decide the rotation order of the subject/sub-field evaluations. They also decide which scientists should be evaluated within a particular sub-field. The PC also appoints a chairman of the Evaluation Committee (EvC) whose task it is to act as reporter of the findings of the EvC to them and to the NFR. This chairman is also one of their own members who is not biased in the sub-field to be evaluated, but familiar with the Swedish academic and research systems and able to assist the international members of the EvC.

Next, a letter is sent to the grant holders by the EvC chairman requesting them to supply names of suitable reviewers. They may suggest as many international evaluators
as they wish, providing these experts fulfill certain conditions. When the composition of
the international EvC is finalized by the PC more detailed instructions and a
questionnaire are sent to the grant holders. The time period to be evaluated is usually the
previous five to six years and a selection of ten reprints may be submitted to
demonstrate the quality of the research projects to be evaluated (Guy-Ohlson, 1997).

For the convenience of both the groups involved an attempt is made to avoid
'rush-hour' times for the writing up of the reports to be submitted for evaluation and for
the site visits. Thus, for example, deadline dates for annual NFR applications are
avoided. In general the pattern is that the grantees have three months to write their
reports, the evaluators have them for approximately three months to review, and then
site visits take place during a one week period, and the final report is published within
the next three to four months. The cycle is then complete and the whole procedure starts
again for the next sub-discipline evaluation.

Though the grant holders suggest names for reviewers, the final choice rests with
the PC. While some stipulated conditions are obvious, the Swedish experience has
shown that certain recommendations are worthy of note when nominating or selecting
reviewers, for example (Guy-Ohlson, 1997):

1. The international experts serving on the EvC must not have
collaborated, nor jointly co-authored papers together

2. The experts must not only cover their own specialty in the field
to be evaluated, but also have a certain breadth of expertise

3. While scientific quality and experience are the prime factors when
inviting experts to participate, it has been found that experts at
the zenith of their career in mid-age and still actively engaged in
building up their own team/department are to be recommended;
they are then eager not only to give but to receive, and alert to
new ideas, approaches and possibly are even interested in
developing contacts after the evaluation

4. It is always necessary to have a large number of names on the
reserve list as refusals for one reason or another are common

5. The number of members serving on the expert panel depends on the number of grant holders within each component part of the sub-field to be evaluated

In addition, the selection and composition of the expert group of evaluators is of utmost importance (Guy-Ohlson, 1997):

...not only must personal chemistry work, but it has also been noted that panel members work best when they are ‘dedicated to the cause.’ By experience it has been found that it works well when they are at an optimal age and position in their own careers (as far as scientific experience is concerned). It is with hesitation that an age is mentioned (late forties/early fifties), but most definitely experts should be at an age when they are still interested in developing their own departments and widening their own fields of interest, and under this should be included willingness to actively read all the papers submitted to them for evaluation and not just look through them on the plane to Sweden. A good deal of energy, enthusiasm and stamina are required to do a good job of evaluation—and the hope is that those who take this assignment will also get something out of it for themselves scientifically (as no one is doing it for the financial remuneration as the honorarium isn’t worth mentioning) (p. 105).

In conducting the evaluations attention is focused on the fact that, though the research projects are the main elements in the evaluation, the EvC is free, and even encouraged, to comment also on structural problems, for example, within the academic system, the age of the doctoral students, and the amounts of money awarded. Under the heading, ‘aspects to be covered by an evaluation’ (Mao, Putnis, Vaughn, & Guy-Ohlson, 1997) the points which should be specifically addressed by the international experts are given in detail. They cover, for example, criteria such as the scientific quality of the results obtained by the grant holders, the scientific value of the proposed projects (including possible improvements by changing the aim and/or direction of the project under evaluation), the merits of the methods used, the capabilities of the project leader and staff, the adequacy of existing and proposed research positions, facilities and
equipment and the question of increased, unchanged or decreased financial support.

Likewise under the heading, 'report of the group' (Mao, Putnis, Vaughn, & Guy-Ohlson, 1997) the expert panel receives instructions as to how they should word their final assessment of the grant holders and their projects. Each grade (excellent, very good, good, fair, and poor) is specifically defined and must be used consequently. "The use of the same standardized terms is necessary for priority selection. It is therefore vital that the grades given and the corresponding terms of recommendation are used consequently throughout an evaluation" (Guy-Ohlson, 1997, p. 104).

Prior to the visit of the expert panel in Sweden, the grantees receive suggested guidelines for the forthcoming site visits. To each head of department or departmental representative a maximum of 20 minutes is allotted for the general presentation of their department to the expert panel. Thereafter a further maximum of 40 minutes is at the disposal of each grant holder for the presentation of their individual research projects and for questioning by the experts. The form of each presentation is left to the grant holder's own discretion. Moreover, "experience has shown that it is absolutely necessary to have a first draft copy of the final evaluation report completed before the experts leave Sweden at the end of their week of site visits. It has proved extremely beneficial and expedient to all to meet each day after the site visits have been completed and after discussion write up the conclusion of the day's work" (Guy-Ohlson, 1997, p. 104).

Five hundred copies of the evaluation report are printed and distributed to the members of the NFR, to the members of the program committees, and to the grant holders. Thereafter the distribution is to the libraries and other scientific funding agencies and to the appropriate governmental departments, but the report may also be obtained on request by other interested parties. These reports are used by the NFR for future planning, setting priorities, and by the PC for the recommendation of awarding
grants. The grant holders themselves cite them in their research application proposals, in their curriculum vitae, and may even make use of them for salary negotiations (Guy-Ohlson, 1997).

United Kingdom

In the following, the United Kingdom’s research structure and context and evaluation model, as well as a brief historical overview of the country is presented.

Historical Overview

As the dominant industrial and maritime power of the 19th century, the United Kingdom of Great Britain and Ireland played a leading role in developing parliamentary democracy and in advancing literature and science. At its zenith, the British Empire stretched over one-fourth of the earth’s surface. The first half of the 20th century saw the United Kingdom’s strength seriously depleted in two World Wars and the Irish republic withdraw from the union. The second half witnessed the dismantling of the Empire and the UK rebuilding itself into a modern and prosperous European nation. As one of five permanent members of the United Nations Security Council, a founding member of NATO, and of the Commonwealth, the United Kingdom pursues a global approach to foreign policy; it currently is weighing the degree of its integration with continental Europe. A member of the EU, it chose to remain outside the Economic and Monetary Union for the time being.

Constitutional reform is also a significant issue in the United Kingdom. The Scottish Parliament, the National Assembly for Wales, and the Northern Ireland Assembly were established in 1999, but the latter suspended due to wrangling over the peace process. In 2006, the United Kingdom had a population of 60.0 million (Central
Research Structure and Context

In the United Kingdom public funding for university-based research is generally provided by a dual-support system (Barker & Lloyd, 1997; Geuna, & Martin, 2003). First, research infrastructure is funded by education departments in the form of block grants paid to each institution. This funding supports a basic level of research activity and is intended to support salaries of permanent staff and facilities, and can be spent at the institution’s discretion. The second line of public research funding comes from the Department of Trade and Industry’s Office of Science and Technology (OST) and is dispersed by the research councils. This stream of funding is intended to support specific research projects and central facilities. A wide range of additional research funding sources include charities, government departments, industry, and through initiatives such as the EU’s FPs (Day, 2004).

With the Education Reform Act of 1988 (House of Commons, 1988) the dual-support funding system underwent dramatic change, and with it created the Universities Funding Council (UFC) and Polytechnics and Colleges Funding Council (PCFC). The Act essentially enabled the UFC and PCFC to become purchasers of academic services, and as a result United Kingdom universities, polytechnics, and colleges were transformed from state-funded public institutions, to client-serving suppliers of research (Geuna & Martin, 2003). In the early 1990s, the UFC and PCFC were combined into a single Higher Education Funding Council (HEFC), and in 1993 separate agencies were created for England, Scotland, Wales, and Northern Ireland. Shortly thereafter, Britain’s polytechnics were granted university status and the government began to encourage competition between the old universities and the former polytechnics.
Evaluation Model

Over the past decade the United Kingdom has developed one of the most comprehensive and sophisticated research evaluation systems in Europe (Hills & Dale, 1995), which has increasingly attracted world-wide interest (von Tunzelmann & Mbula, 2003). The RAEs are periodic evaluations of the research undertaken by United Kingdom universities, initiated by the HEFCs to inform the allocation of research funding. The RAE is designed to give each university research unit a quality rating and reward excellence. From the start, the University Grants Committee (UGC) saw the RAE as a basis for informing selective allocation of block grants paid to higher education institutions for research. Funding formulas are based on RAE quality ratings and units receiving the highest ratings are given greater weight in the distribution of approximately £1 billion ($1.8 billion USD) per year of public funding. This mechanism aims to ensure that the best research in the United Kingdom is protected and developed (Wooding & Grant, 2003). The costs of the RAE, including opportunity costs, have been variously estimated at between £27 million ($47 million USD) and £37 million ($64 million USD) per exercise (Roberts, 2003); less than 0.8% of the total funds distributed on the basis of the exercise.

The first RAE was conducted in 1986 and has been repeated in 1989, 1992, 1996, and 2001, with the next RAE scheduled for 2008 (RAE 2008, 2005). The 2001 RAE was performed jointly by the Higher Education Council for England (HEFCE), the Scottish Higher Education Funding Council (SHEFC), the Higher Education Funding Council for Wales (HEFCW), and the Department of Education for Northern Ireland (DENI). The RAE does not distinguish between basic or applied research, and in

---

282 The United Kingdom's RAE has been influential nationally and, as a technique, internationally. However, the results for the less successful hang around their necks for years (Nature, 2005).

283 The acronym HEFC is used here to denote these four councils.
2001 the operational definition adopted for research was:

...original investigation undertaken in order to gain knowledge and understanding. It includes work of direct relevance to the needs of commerce and industry, as well as to the public and voluntary sectors; scholarship; the invention and generation of ideas, images, performances and artifacts including design, where these lead to new or substantially improved insights; and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction. It excludes routine testing and analysis of materials, components and processes, for example, for the maintenance of national standards, as distinct from the development of new analytical techniques (HERO, 1999a).

The RAE has been characterized as an ex post evaluation system (Geuna & Martin, 2003) governed by (a) clarity, (b) consistency, (c) continuity, (d) credibility, (e) efficiency, (f) neutrality, (g) parity, and (h) transparency (HERO, 1999a). All research activities within a university are categorized into units of assessment (UoA). In 2001, 68 UoAs were defined under five Umbrella Groups (I. Medical and Biological Sciences [UoA 1-17]; II. Physical Sciences and Engineering [UoA 18-32]; III. Social Sciences [UoA 33-44, 68, 69]; IV. Area Studies and Languages [UoA 45-56]; and V. Arts and Humanities [UoA 57-67]). The role of the Umbrella Groups is to facilitate cross-panel consistency in the assessment process and application of standards (a-g above) with particular attention given to areas of work which span the boundaries of UoAs, including interdisciplinary research. Some researchers (Loder, 1999; Tait, 1999) have argued that the RAE unfairly penalizes interdisciplinary research and therefore the HEFC has encouraged departments and groups to submit their work to the “most appropriate panel, and to suggest second panels to consider submissions in parallel” (Geuna & Martin, 2003, p. 281).

For each UoA a panel of 10 to 15 experts is assembled using a process by which professional associations and learned societies nominate panel candidates. Panel Chairs,
appointed jointly by HEFCE, SHEFC, HEFCW, and DENI as a result of being nominated by prior RAE panel members from among their own number, select panel members from these nominees on the basis of experience, standing, and representation of user communities (Geuna & Martin, 2003; HERO, 1999b). In 2001, more than 1,300 professional associations and learned societies nominated candidates for these panels, and 3,024 twenty-four nominations were received from over 400. In 2001, the RAE was undertaken by 60 assessment panels (although there are sixty-eight UoAs, some are grouped into Joint Panels [e.g., UoAs 5-8: pre-clinical studies, anatomy, physiology and pharmacology are grouped into a single Joint Panell] consisting of 685 members. Overall, 20% of the 2001 RAE membership were women and 10% originated from post-1992 institutions. The geographical distribution of panel membership for the 2001 RAE was generally consistent with the relative sizes of the United Kingdom territories (England = 78%, Scotland = 13%, Wales = 5%, Northern Ireland = 2%), with members drawn from more than 100 institutions.

Every department or group within a university is assigned to a UoA, and hence to a panel. Each of these departments or groups is required to submit a standardized submission of information on research performance to the appropriate UoA, including the following (HERO, 1999c):

1. **Overall staff summary** (RA0). Summary information on all academic staff and academic support staff in each submitting institution

2. **Research active individuals details** (RA1). Detailed information on individuals selected by the institution for inclusion as research active

3. **Research output** (RA2). For each individual named as research active up to four items of research output produced during the period 1 January 1994 to 31 December 2000 for arts and humanities subjects (UoAs 45 to 67 inclusive); and 1 January 1996 to 31 December 2000 for other subjects (UoAs 1 to 44, 68 and 69)
4. Research students (RA3a). Numbers of full-time and part-time postgraduate research students and degrees awarded

5. Research studentships (RA3b). Numbers of postgraduate research studentships and source of funding

6. External research income (RA4). Amounts and sources of external funding

7. Textual description (RA5). Including information about the environment, structure, policies and strategies within which research is undertaken and developed

8. General observations and additional information (RA6)

9. and...information about indicators of research excellence and peer esteem which cannot be given elsewhere

Methodologically, the RAE is broadly based on an informed peer-review process.

In 2001, the RAE criteria and working methods were published over a year before submissions were due. For example, the quality assessment criteria\(^{284}\) for the Psychology Panel (UoA 13) were (HERO, 1999d):

1. The quality of publications and other forms of research output cited

2. Research student activity

3. External research income as an indicator of esteem

4. Evidence of the research environment and infrastructure of the submission, its vitality and prospects for continuing development; the distinction of its members and the impact of their work

Using submitted information (RA0-RA6) and the quality assessment criteria (a-d), panels judge the quality of each department or group's research and assign a quality rating ranging from 1 to 5\(^*\) (HERO, 1999a). As shown below in Table 10, the ratings are determined by the amount of work judged as attaining 'national' and 'international' levels

---

\(^{284}\) The quality of publications and other forms of research output cited carried the most weight in the assessment process, "but all indicators...[are]...scrutinized and...contribute to the overall assessment of a submission" (HERO, 1999d).
of excellence. In 2001, the RAE panels evaluated more than 2,400 submissions and 150,000 publications.

**Table 10**

United Kingdom’s Research Assessment Exercise Rating Scale

<table>
<thead>
<tr>
<th>Rating scale</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>5*</td>
<td>Quality that equates to attainable levels of international excellence in more than half of the research activity submitted and attainable levels of national excellence in the remainder.</td>
</tr>
<tr>
<td>5</td>
<td>Quality that equates to attainable levels of international excellence in up to half of the research activity submitted and to attainable levels of national excellence in virtually all of the remainder.</td>
</tr>
<tr>
<td>4</td>
<td>Quality that equates to attainable levels of national excellence in virtually all of the research activity submitted, showing some evidence of international excellence.</td>
</tr>
<tr>
<td>3a</td>
<td>Quality that equates to attainable levels of national excellence in over two-thirds of the research activity submitted, possibly showing evidence of international excellence.</td>
</tr>
<tr>
<td>3b</td>
<td>Quality that equates to attainable levels of national excellence in more than half of the research activity submitted.</td>
</tr>
<tr>
<td>2</td>
<td>Quality that equates to attainable levels of national excellence in up to half of the research activity submitted.</td>
</tr>
<tr>
<td>1</td>
<td>Quality that equates to attainable levels of national excellence in none, or virtually none, of the research activity submitted.</td>
</tr>
</tbody>
</table>

For research funds the volume and quality of research is decisive. Sums of money are made available within each of the UOAs. The amount of funds allocated to each institution within each subject is proportional to a volume measure multiplied by a quality measure as shown in Equation 3, where:

$$\text{Amount} = \text{Quality} \times \text{Volume} \quad (3)$$

The research ratings are converted into a funding scale, ranging from 0 to 4.05. Ratings 1 and 2 receive no funding, while a rating of 5 receives four times as much funding as a rating of 3b for the same volume of research activity. In order to devise a single overall unit measure, the volume of research is measured in each UOA using five
separate components weighted as follows in Equation 4:

\[
\text{Research active academic staff} = 1 \times \text{number of full-time equivalent research staff} \quad (4a)
\]

\[
\text{Research assistants} = 0.1 \times \text{number of FTE research assistants} \quad (4b)
\]

\[
\text{Research fellows} = 0.1 \times \text{number of FTE research fellows} \quad (4c)
\]

\[
\text{Postgraduate research students} = 0.15 \times \text{number of postgraduate research students} \quad (4d)
\]

\[
\text{Research income from charities} = \frac{0.25}{25,000} \times \text{average of two years’ income from charities} \quad (4e)
\]

After the 2001 RAE, the HE.PCs were unable to provide the estimated £170 million ($295 million USD) of additional funding required to reward excellence in research as revealed by the exercise. In part, this was due to the increase in the number of departments and groups scoring greater than 5 in 2001 as compared with prior RAES. In 1996, 43% of United Kingdom research was rated as meeting or exceeding national or international standards, whereas in 2001 63% meet or exceeded these standards. This, and outcries from academia and elsewhere (e.g., Davis, 2002; Farrar, 2002) led to a review of the system. In 2002, a steering committee lead by Sir Gareth Roberts was appointed to lead the review and charged with investigating different approaches to the definition and evaluation of research quality. His review included soliciting submissions from interested parties, workshops (Wooding and Grant, 2003) and meetings, an operational overview of the 2001 RAE, and von Tunzelmann and Mbula’s (2003) study of international approaches to research assessment. Roberts delivered his final report to the United Kingdom funding councils in May 2003, and recommended:

...a radical overhaul of the Research Assessment Exercise (RAE). They do not however represent a wholesale rejection of the RAE and the principles upon which it was built. All who examine the impact of the RAE upon UK research and its international reputation must, I think,
agree that it has made us more focused, more self-critical and more respected across the world...I have proposed a system which appears more complex than what has gone before. However, I acknowledge a sense in which these proposals do sacrifice simplicity for efficiency and fairness (Roberts, 2003, p. 2).

In his report Sir Roberts made sixteen recommendations, which if implemented would have required a complete restructuring of the current system. For example, Roberts’ recommendation of giving performance metrics (e.g., bibliometrics, grant income) an increased role in the RAE found little support in the funding bodies, and the RAE will continue to be based on peer review, although panels are “encouraged to use metrics to support decisions where appropriate” (Day, 2004, p. 20). However, Roberts’ recommendation of replacing the RAE rating scale with a continuously graded quality profile, with four starred levels, was accepted.\footnote{For each submission the panel would produce a ‘quality profile’ (Roberts, 2003). For each submission the panel would decide how much work could be defined as meriting one, two or three stars (or no stars).} This system would emphasize that the focus of the assessment is to discriminate between very good research and the very best. It also provides for what is, in effect, a continuous grading scale, eliminating the problems created by grade boundaries. This recommendation was based on Roberts’ observation of a ceiling effect where, for example, 80% of the researchers whose work was assessed received one of the three top grades (4, 5, and 5*), while 55% received one of the top two grades (5 and 5*) in 2001. Therefore, he argued that the amount of discrimination provided by the exercise is therefore less than the length of the rating scale would suggest.

United States

In the following, the United States’ research structure and context and evaluation model, as well as a brief historical overview of the country is presented.
Historical Overview

Britain's American colonies broke with the mother country in 1776 and were recognized as the new nation of the United States of America following the Treaty of Paris in 1783. During the 19th and 20th centuries, thirty-seven new states were added to the original thirteen as the nation expanded across the North American continent and acquired a number of overseas possessions. The two most traumatic experiences in the nation's history were the Civil War (1861-1865) and the Great Depression of the 1930s. Buoyed by victories in World Wars I and II and the end of the Cold War in 1991, the United States remains the world's most powerful nation state. The economy is marked by steady growth, low unemployment and inflation, and rapid advances in technology. The United States had a population of nearly 300.0 million in 2006 (Central Intelligence Agency, 2006).

Research Structure and Context

Research in the United States faces quite different assessment problems than other nations. First, it is larger in sheer magnitude and produces nearly one-third of the world's research output (Ab Iorwerth, 2005). Second, there is no system of national universities in the United States, "so questions of evaluation of university units, which constitute at least half the problem in many other countries, do not arise at [the] national level" (Cozzens, 2000, p. 6). Third, unlike many other nations, there are no research councils in the United States, but rather a relatively small NSF which shares the responsibility of university research with numerous other mission-oriented agencies (Cozzens, 2000). The mechanism by which the United States allocates resources for research is the federal budget process, which is "neither rational nor systematic...its major strength, however, is that it works and...has helped to foster a uniquely vigorous
and creative research enterprise” (Teich, 1997, p. 9).

Perhaps because of worries about increased government intervention, the important notion of research as an autonomous pursuit, free of interference by sponsors, was asserted at this time, one of the major influences being the United States’ Presidential Science Advisor, Vannevar Bush (1945/1960). Bush managed to instill the idea of a generously funded yet self-governing scientific establishment. By stressing the importance and inevitable benefits of research, he helped legitimate the ‘linear model’ where inputs in research would eventually feed into technological innovation. Bush argued that it would be self-defeating to attempt to constrain the creativity of research, and that science was most fertile if it was not under direct governmental control (National Science Board, 2000). In the last 50 years, the rationale for government support of research has been the contribution of science and technology to military security and national prestige in the Cold War environment, coupled with a sense—taken “mainly on faith” (Teich, 1997, p. 9)—that a strong research community will more than pay for itself in economic and social benefits.

Following the end of the Cold War, old questions about the control of the United States’ research agenda (Weinberg, 1989) and procedures and methods for determining the allocation of resources among fields and disciplines, research institutions, and regions (Cozzens, 2000) were once again surfacing. In response to increasing demands for accountability, the Planning-Programming-Budgeting System (PPBS) was introduced in the mid-1960s, Management By Objective (MBO) in the late 1960s, Zero-Based Budgeting (ZBB) in the 1970s, and Total Quality Management (TQM) in the 1980s and 1990s (Roessner, 2002).

Unlike most parts of the world, the United States Congress explicitly adopted consumer-based financial support in lieu of direct grants to institutions for research. This
policy was designed in part upon the belief that student-based funding would create greater innovation and quality in academic programs by stimulating a more competitive academic market. Similarly, federal support for research was allocated to individual researchers and teams through a competitive grant system rather than through institutional grants (Dill, 2003, March). This market-oriented system, in which research allocations are made not to individual institutions, but to individual researchers on an open, competitive basis (e.g., NSF\textsuperscript{286} and NIH\textsuperscript{287} proposal evaluations, which are characterized in part by their predictive and sorting function), means that the total research funding of any institution represents the accumulation of these individual, competitive market transactions, rather than political or peer judgments about unit or university quality (Dill, 2001). The federal grants system does depend upon peer judgments of research proposals, however, rather than measures of research output.

The mechanism through which the United States allocates resources for basic research at the macro-level is the federal budget process. This process is neither rational nor systematic. It is, indeed, unsystematic, confusing, and in many respects, irrational (Teich, 1997, p. 9).

**Evaluation Model**

During the 1960s, two research evaluations were conducted which are considered milestones in the evaluation of United States’ research; the 1969 evaluations of the State Technical Services (STS) Program and the DOD’s Project Hindsight, conducted by Arthur D. Little, Inc. (ADL). In the STS evaluation ADL estimated the economic effects of state’s STS technology transfer activities by converting estimates of successful project’s results into estimates of increased sales revenues, calculations of benefits, and

\textsuperscript{284} NSF applies two explicit criteria in evaluating proposals: (i) intellectual merit and (ii) broader impacts (National Science Foundation, 2004).

\textsuperscript{287} NIH applies five explicit criteria in evaluating proposals: (i) significance; (ii) approach; (iii) innovation; (iv) investigators; and (v) environment (National Institutes for Health, 2004).
aggregated effects of the program (Roessner, 2002). The evaluation of Project Hindsight, however, was an attempt to develop an understanding of the costs and benefits of United States support for basic versus applied research (Sherwin & Isenson, 1967), and focused on twenty major weapons systems supported by the DOD and traced the development of each backwards in time twenty years in an attempt to identify the research outputs that contributed to their development.

The evaluation of United States’ research in the 1970s and 1980s was characterized by a series of studies that sought to assess the economic returns from federal investments in research and social rate of return (Roessner, 2002). NASA led the way with Mathematica, Inc. (1976) and Mathtec, Inc.’s (1977) ground-breaking studies which attempted to measure the economic benefits of NASA R&D. The social rate of return studies (e.g., Mansfield, 1980; Mansfield, Rapoport, Romeo, Wagner, & Beardsley, 1977; Tewksbury, 1980), on the other hand, sought to estimate the social benefits that accrue from changes in technology, and “relate the value of these benefits to the cost of the research underlying the technological changes” (Roessner, 2002, p. 87). Like the NASA studies, the social rates of return studies were subjected to severe methodological criticisms and discontinued (Georghiou & Roessner, 2000; Hertzfeld, 1992; Roessner, 2002). At the same time, evaluation offices were established by both NSF and NIH in the mid-1970s, both of which pioneered unique methods for evaluating their projects and programs. The NIH central evaluation office, for example, established one of the earliest databases of publications for bibliometric analyses, begun by the RAND Corporation in 1974. These bibliometric studies were used to report on the quantity and character of publications, and the results were compared with peer evaluations.

Simultaneously, the NSF evaluation office was conducting studies of peer review systems. Like the NIH, NSF also considered bibliometric indicators as a relevant
approach for evaluating research, and in 1976 published *Evaluative Bibliometrics* (Narin, 1976). More importantly, NSF commissioned its first documented evaluation of research which employed both peer review and bibliometric methods of assessing research quality in its evaluation of Materials Research Centers (Cozzens, 2000). However, the use of bibliometric techniques has declined substantially since the early 1990s, and the techniques have not been embraced by other United States agencies to a significant extent (Roessner, 2002).

In the early 1990s, the United States Congress set aside funds for serious evaluation of NSF’s science education programs, and an office of professional evaluators was established in the Education and Human Resources (I) Directorate. Unfortunately, to demonstrate the general worth of their programs rather than to make changes in program resources or focus, virtually all of the United States agencies collected and published “success stories, highlighting breakthroughs or high-impact research areas in their laboratories or grantees” (Cozzens, 2000, p. 6).

Simultaneously, in two separate divisions of the National Institute of Standards and Technology (NIST), plans were created to integrate multiple evaluations of the Advanced Technology Program (ATP) and the Manufacturing Extension Partnership (MEP), both of which were newly created federal programs of research, technology development, and technical assistance and transfer. These evaluation commitments were significant because they “represent substantial programmatic commitments to research evaluation rather than ad hoc or periodic, stand-alone evaluation efforts, the norm for research evaluation in the USA” (Roessner, 2002, p. 88). For example, analyses and evaluations were conducted on almost every aspect of ATP (Ruegg, 1998), including economic analyses reminiscent of those performed in the 1970s and 1980s. The ATP evaluation represents one of the most sophisticated and thorough research evaluations
ever conducted in the United States, and from more than forty-five studies commissioned by ATP between 1990 and 2000. *A Toolkit for Evaluating Public R&D Investment* (Ruegg & Feller, 2002) was produced. ATP was also the focus of a summative evaluation conducted by the GAO (GAO, 1995), commissioned by the United States Congress. However, the GAO report was seen as inconclusive by advocates and critics alike as to whether or not the program was worth the public investment.

Passage of the GPRA in 1993, however, “reflected a desire on the part of the public and their representatives in Washington for more effective and efficient use of public funds” (COSEPUP, 2001, p. 7), including research (COSEPUP, 1999, 2001; Cozzens, 2000; Roessner, 2002). The GPRA was developed by analysts with extensive experience in program evaluation, intended to supplement formal program evaluation with short-term performance monitoring (Cozzens, 2000), and required each agency to submit three documents to the United States Congress and OMB, which were to include: (1) a strategic plan, which was to cover all agency functions over a five-year period and be updated every three years; (2) a performance plan, in which performance levels were quantified; and (3) performance reports which provided information on actual versus projected performance. Performance plans and reports were to be submitted annually.

GPRA differed from previous systems in its legislative foundations, which potentially gave it greater authority and staying power (Roessner, 2002). Coinciding with the introduction of the GPRA, United States agencies that supported research began to develop mechanisms for assessing their agendas and programs of research. Foremost among these were the NSF and the NIH with their occasional program evaluations, followed by the mission-oriented agencies such as the DOE and the Office of Naval Research (ONR), which already had systems of program review in place (Cozzens, 2000).
In 1999, COSEPUP attempted to address the evaluation of federal research programs in response to GPRA in its report *Evaluating Federal Research Programs: Research and the Government Performance and Results Act*. That report, and 2001's *Implementing the Government Performance and Results Act for Research* (COSEPUP, 2001) indicated that federal research programs were best evaluated using expert review on the basis of three criteria:

1. **Quality.** Review of the quality of research via peer review is the most common form of expert review; the sine qua non of quality review is objectivity.

2. **Relevance.** The extent to which agency's research addresses subjects in which new understanding could be important in fulfilling the agency’s mission.

3. **Leadership.** Extent to which agency's research is at the forefront of a field internationally.

Expert review, as opposed to traditional peer review, includes users of research, whether they are in industry, non-government organizations, or public health organizations or other members of the public who can evaluate the relevance of the research to agency goals (COSEPUP, 2001). Despite the call for expert review, peer review remains the “backdrop against which all other types of research evaluation appear, and often the standard against which their validity is judged” (Roessner, 2002, p. 86) in the United States. Applied to ex ante, or a priori, project selection, peer review has remained the dominant method by which proposals for research are rated, and has also

---

288 Of the dozen or more agencies in the United States that support (large-scale) research, or research programs, five participated in the 2001 COSEPUP study: the Department of Defense (DOD), Department of Energy (DOE), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), and National Science Foundation (NSF).

289 The COSEPUP (2001) report does not provide a specific definition for the quality criteria, but rather only a suggested method for determining quality.

290 The rationale is that the United States must be performing at the forefront of a field if it is to understand, appropriate, and capitalize on current advances in the field (COSEPUP, 1993). This criterion “review can be accomplished by the technique if international benchmarking; an exercise carried out by a panel of non-US experts and US experts whose technical expertise and international perspective qualify them to assess the standing of a research program or an entire field” (COSEPUP, 2001, p. 15).
been the method embraced by the scientific community for ex post, or a posteori, assessment of research quality in the United States.

Since 2002, GPRA has been combined with federal R&D investment criteria and PART. Whereas GPRA establishes a broad statutory framework for management and accountability, the R&D investment criteria and PART are focused more on simplified measures of performance for budget decisions.

The United States OMB R&D Investment Criteria (U. S. Office of Management and Budget, 2003) are:

1. Relevance
   a. Programs must have complete plans, with clear guidelines and priorities
   b. Program must articulate the potential public benefits of the program
   c. Programs must document their relevance to specific presidential priorities to receive special consideration
   d. Program relevance to the needs of the nation, of fields of science and technology, and of program “customers” must be assessed periodically through retrospective external review

2. Quality
   a. Programs allocated funds through means other than a competitive, merit-based process must justify funding methods and document how quality is maintained
   b. Program quality must be assessed periodically through retrospective expert review

3. Performance
   a. Programs may be required to track and report relevant program inputs annually
   b. Programs must define appropriate output and outcome measures, schedules, and decision points
The R&D PART was developed to:

...assess and improve program performance so that the Federal government can achieve better results. A PART review helps identify a program’s strengths and weaknesses to inform funding and management decisions aimed at making the program more effective. The PART therefore looks at all factors that affect and reflect program performance including program purpose and design; performance measurement, evaluations, and strategic planning; program management; and program results. Because the PART includes a consistent series of analytical questions, it allows programs to show improvements over time, and allows comparisons between similar programs (U.S. Office of Management and Budget, 2006b).

The components and weightings of R&D PART are:

1. Program purpose and design (20% weighting)—Questions address R&D investment criteria of program relevance
   
   a. Is the program purpose clear?
   
   b. Does the program address a specific and existing problem, interest, or need?
   
   c. Is the program designed so that it is not redundant or duplicative of any other federal, state, local, or private effort?
   
   d. Is the program design free of major flaws that would limit the program’s effectiveness or efficiency?
   
   e. Is the program design effectively targeted so that resources will address the program’s purpose directly and will reach intended beneficiaries?

2. Strategic planning (10% weighting)—Questions address prospective aspects of the R&D investment criteria

   a. Does the program have a limited number of specific long-term performance measures that focus on outcomes and meaningfully reflect the purpose of the program?

   b. Does the program have ambitious targets and time frames for its long-term measures?

   c. Does the program have a limited number of specific annual performance measures that can demonstrate progress toward
achieving the program's long-term goals?

d. Does the program have baselines and ambitious targets for its annual measures?

e. Do all partners (including grantees, subgrantees, contractors, cost-sharing partners, and other government partners) commit to and work toward the annual and/or long-term goals of the program?

f. Are independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need?

g. Are budget requests explicitly tied to accomplishment of the annual and long-term performance goals, and are the resource needs presented in a complete and transparent manner in the program's budget?

h. Has the program taken meaningful steps to correct its strategic planning deficiencies?

3. Additional questions for R&D programs

a. If applicable, does the program assess and compare the potential benefits of efforts within the program and (if relevant) to other efforts or similar programs that have similar goals?

b. Does the program use a prioritization process to guide budget requests and funding decisions?

4. Program management (20% weighting)—Questions address prospective aspects of quality and performance in the R&D investment criteria, as well as general program management issues

a. Does the agency regularly collect timely and credible performance information, including information from key program partners, and use it to manage the program and improve performance?

b. Are federal managers and program partners (including grantees, subgrantees, contractors, cost-sharing partners, and other government partners) held accountable for cost, schedule, and performance results?

c. Are funds (federal and partners) obligated in a timely manner and
spent for intended purposes?

d. Does the program have procedures (e.g., competitive sourcing or cost comparisons, information technology improvements, appropriate incentives) to measure and achieve efficiencies and cost-effectiveness in program execution?

e. Does the program collaborate and coordinate effectively with related programs?

f. Does the program use strong financial management practices?

g. Has the program taken meaningful steps to address its management deficiencies?

5. Additional questions for R&D programs

a. For R&D programs other than competitive grants, does the program allocate funds and use management processes that maintain program quality?

6. Program results and accountability (50% weighting)—Questions address retrospective aspects of the R&D investment criteria, with emphasis on performance

a. Has the program demonstrated adequate progress in achieving its long-term performance goals?

b. Does the program (including program partners) achieve its annual performance goals?

c. Does the program demonstrate improved efficiencies of cost-effectiveness in achieving program goals each year?

d. Does the performance of this program compare favorably to other programs, including government, private, etc. with similar purposes and goals?

e. Do independent evaluations of sufficient scope and quality indicate that the program is effective in achieving its goals?

As shown in Table 11, the current distribution for the 793 programs that have already been through a PART review shows that approximately a quarter of them received a rating of Results Not Demonstrated. Fifteen percent were determined to be
Effective. Approximately 30% were found to be Moderately Effective or Adequate. Only 4% have been determined to be Ineffective (American Evaluation Association, 2006c).

Table 11

PART Results to Date

<table>
<thead>
<tr>
<th>Rating</th>
<th>N</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective</td>
<td>124</td>
<td>15%</td>
</tr>
<tr>
<td>Moderately effective</td>
<td>231</td>
<td>29%</td>
</tr>
<tr>
<td>Adequate</td>
<td>219</td>
<td>28%</td>
</tr>
<tr>
<td>Ineffective</td>
<td>28</td>
<td>4%</td>
</tr>
<tr>
<td>Results not demonstrated</td>
<td>191</td>
<td>24%</td>
</tr>
</tbody>
</table>

Although the fraction of the United States budget invested in research is relatively small, it is highly visible. In 1998, federal funds supported some $20.2 billion worth of basic research, almost half of which went to the NIH (COSEPUP, 2001). In the same year, almost $50 billion more was spent on applied research, of which a large portion was devoted to “the procurement and testing of weapons systems” (COSEPUP, 2001, p. 7). In 2004, federal support for university research alone was estimated at more than $21.8 billion, an increase of nearly 50% over 1997 (Tash & Sacks, 2004). However, unlike many other nations, the United States has yet to develop wide-scale systems for evaluating its research on a national level.

**Fundamental Characteristics of International Research Evaluation Models**

The remainder of this chapter is intended to supplement the historical overviews, the research structures and contexts, and the evaluation models for each of the 16 countries. Herein, the fundamental characteristics of these research evaluation models
are presented in terms of their: (i) primary reasons and motives; (ii) basic units of assessment; (iii) core methods; (iv) key indicators and criteria; (v) systemization and consistency (i.e., general evaluation strategy); and (vi) funding model archetype.

Primary Reasons and Motives

Analysis of the 16 national systems sampled revealed that most countries reasons and motives for evaluating their government-financed research closely approximate the reasons and motives described in Chapter I (i.e., accountability and efficiency, resource allocation, improvement, and decision making). As shown in Table 12, 94% of the national systems evaluate their publicly-funded research for reasons of accountability and efficiency, 63% for resource allocation, 50% for improvement, and 31% for other types of decision making (e.g., setting research policies or priorities). A large majority (81%) evaluate their publicly-funded research for two or three of these reasons.

Again, as described in Chapter I, as a purpose for evaluating research, particularly publicly-funded research, accountability and efficiency is the responsibility for the justification of expenditures, decisions, or the results of research efforts. Resource allocation involves matters such as national priority setting, which normally includes the distribution of research funding (Coryn, 2007, January; Scriven, 2006f). Normally, improvement is a secondary function of national-level evaluation of publicly-funded research, expected to occur as a result of competition for research monies. Decision making, as described here, involves matters such as selection, prioritization, and prediction (see Chapter I).
Table 12

International Research Evaluation Models’ Primary Reasons and Motives for Evaluating Research

<table>
<thead>
<tr>
<th>Reason</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability and efficiency</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Resource allocation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improvement</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Decision making</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

262
Basic Units of Assessment

Typically, the national research evaluation models emphasize one or more of the following eight units of assessment:

Research products. Research products are normally confined to scholarly publications, but may also include patents, computer programs, and other technologies and innovations.

Individual researchers. Individual researcher's performance; usually includes research products.

Research groups. Researchers from different institutions or universities active in the same specialty or discipline.

Programs or projects. Programs and projects usually in relation to national priority areas (e.g., renewable energy research); includes large- and small-scale government-financed research programs and projects.

Departments. Departments are usually discipline-specific units (e.g., chemistry, education, physics, mathematics, psychology, sociology) within an institution.

Institutions. In most countries, institutions are typically places of higher learning/education (i.e., universities).

Disciplines. Entire scientific disciplines or research collectives.

Policies. National research or research evaluation policies; including research funding policies.

As shown in Table 13, the most common unit of assessment in the sampled countries is the institution (69%), followed by research products (44%). Only the United Kingdom uses departments as a unit of assessment; albeit, within institutions via assessment of research products.
<table>
<thead>
<tr>
<th>Unit of Assessment</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research products</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Individual researchers</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Research groups</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Programs or projects</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Departments</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Institutions</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Disciplines</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Policies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Core Methods

Methodologically, most national systems typically use one or more of the following 13 approaches or strategies to evaluate their publicly-financed research:

**Bibliometrics.** Typically, bibliometric methods are confined to scholarly publications (including patents) and citations to them; it also includes spatial mapping, data mining, data visualization (e.g., research networks), webometrics, and similar techniques

**Case studies.** Gathering and analyzing data about one or a small number of examples as a way of studying a broader phenomenon; done on the assumption that the example (i.e., case) is in some way typical of the broader phenomenon

**Comparative studies.** Studies of more than one event, group, or nation to isolate factors that explain patterns; most often cross-national comparisons

**Cost analysis.** Most often, classical costs-benefits, cost-effectiveness, cost-utility, cost-feasibility, return on investment analyses, and financial ratio analyses; rarely considers non-monetary and other types of costs

**Expert panels (internal).** Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); internal expert panels consist only of experts within the country/nation

**Expert panels (external).** Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); external expert panels consist only of experts outside the country/nation
**Expert panels (mixed).** Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); mixed expert panels consists of both internal and external experts

**Interviews.** A conversation between two or more people where questions are asked by the interviewer to obtain information from the interviewee; Interviews can be divided into two general types, interviews of assessment and interviews for information

**Observations.** Observations are usually conducted by auditors or expert panels; observers do not normally interact with those being observed; usage varies; often a supplement to other methods

**Self-evaluations.** Evaluating and reporting on the quality or value of one's own work; often a supplement to other methods

**Site visits.** Site visits are usually conducted by auditors or expert panels; unlike observations, observers interact with those being observed; usage varies; often a supplement to other methods

**Strategic plans.** Analysis of an individual’s, group’s, project or program’s, or institution’s strategic research plans; sometimes used to set performance targets or standards; often a supplement to other methods

**Surveys.** Sampling from a population in order to make inferences about the population; usually in the form of questionnaires, less often in the form of interviews; sometimes a census of an entire population; usage varies; often a supplement to other methods

The most commonly employed methodology is the expert panel (see Table 14). Every country in the sample uses at least one of the varieties of expert peer panels; 31% using primarily internal peers; 19% using primarily external peers; and 50% using primarily mixed-peer panels. Other widely used approaches include network analysis, tracer methodologies, spillover analysis, and data mining and visualization techniques (Ruegg & Feller, 2002; Ruegg & Jordan, 2007). Not surprisingly, the focus of the current debate has been mostly on what method to use (Julnes & Rog, 2007).
<table>
<thead>
<tr>
<th>Method</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliometrics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Case studies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Comparative studies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cost analyses</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Expert panels (internal)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Expert panels (external)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Expert panels (mixed)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Interviews</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-evaluations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Site visits</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Strategic plans</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Surveying</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 14
International Research Evaluation Models' Core Methods
Key Indicators and Criteria

The typical quality indicators and criteria used by most national systems include one or more of the following:291

*Patents.* Patent applications and patents granted by EPO, USPTO, and JPO; frequently viewed as indicators of innovation

*Local, regional, national, and international impact.* Impact of research; normally estimated using bibliometric techniques, expert panel assessment

*Researchers.* Professionals engaged in the conception or creation of new knowledge, products, processes, and in the management of research

*Students.* Students enrolled in research-related programs; sometimes students enrolled in any program of study

*Degrees awarded.* Students completing research-related programs of study; usually at the doctoral level

*External research funding.* Funding received from non-governmental sources (e.g., private sector)

*Esteem.* Awards, keynote speeches and addresses, and journal editorships

*Research inputs.* Equipment, staff, funding, and other relevant inputs

*Research outputs.* All varieties of research outputs, including, but not limited to scholarly publication, products, and patents

*Research process.* Everything that occurs prior to research outputs

By far, most national systems place the greatest emphasis on the impacts of research (see Table 15); in particular international impact (100%). The way in which these impacts are estimated, however, varies widely (e.g., bibliometrics, peer judgment). Research outputs are also commonly used as quality indicators (by 81% of the sampled countries); yet, sometimes in reference to quantity rather than quality.

291 Economic indicators, such as GERD, BERD, and GBAORD, have not been included here as most countries typically monitor these data for policy decisions regarding research expenditures.
Table 15

International Research Evaluation Models’ Key Indicators and Criteria

<table>
<thead>
<tr>
<th></th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Local impact</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regional or national impact</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>International impact</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Researchers</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Students</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Degrees awarded</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>External research funding</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Esteem</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Research inputs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Research outputs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Research process</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Systemization and Consistency

Most national-level research evaluation models can be classified in terms of systemization and consistency into terms of two general types, which are (Campbell, 2002):

- **Type A**
  - Type A research evaluation systems apply an approach which is systemic and consistent.

- **Type B**
  - Type B research evaluation systems use pluralized approaches, and can be characterized by a high degree of situation-specific variability in terms of their conceptions and methods.

Type A models use a system, method, or approach which is systematic and consistent. That is, the criteria brought to bear, the standards applied, and the methodological approach does not vary to any great degree across time. Researchers evaluated under Type A models generally know what to expect from the system and what is expected of them. By contrast, Type B systems are often haphazard and highly variable. Very often these types are experimental as governments try to discern relevant criteria and standards, as well as efficient and effective methods for evaluating their government-financed research.

As shown in Table 16, 37% (6 of 16) were classified as Type A systems versus 63% (10 of 16) being classified as Type B systems. However, many of these national systems are considered experimental, being reformed, or currently under development, making them difficult to correctly classify. In such cases, these models were placed in the Type B category as they cannot be considered either systematic or consistent.
Table 16
International Research Evaluation Models Systemization and Consistency

<table>
<thead>
<tr>
<th></th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Type B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Funding System Archetypes

Another useful way to conceptualize and classify the various international systems is by their research funding system model, or archetype, of which there are three major categories. Although most systems have elements of all types, it is possible to classify the systems according to three major types of mechanisms for evaluating publicly-financed research, particularly as they relate to allocating research funding:

- **Type I**: Large-scale performance exercises of various hues; future funding allocations are made on the basis of prior performance; sometimes used in conjunction with Type II and III models.

- **Type II**: Bulk funding models; generally block grant allocations of research funds; sometimes a mix of direct funding for public research institutions and universities and competitive grants programs offered by independent funding agencies.

- **Type III**: Indicator-driven mechanisms; research financing is distributed on the basis of student numbers, external funding, teaching volume, and other quantifiable measures via various funding formulas.

Not considered in this classification, however, is the centralized versus decentralized, or mixed systems for funding research (Conraths & Smidt, 2005; OECD, 2003a). Most countries have centralized research funding mechanisms (i.e., research funding comes from one government agency). Belgium and the United States, however, are decentralized in that multiple agencies or government branches fund a large portion of the countries’ research. In any case, 31% (5 of 16) were classified as Type I models, 44% (7 of 16) as Type II models, and the remaining 25% (4 of 16) as Type III models (see Table 17). Although the Netherlands’ model was classified as Type I, this exercise presently has no connection with the level of funding received, but is in force to improve the public accountability of research activity.
Table 17
International Funding System Archetypes

<table>
<thead>
<tr>
<th>Type</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Type II</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Type III</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Conclusions

In most countries, the competition for government research monies is getting increasingly competitive, which is particularly evident in systems that operate on performance-based funding (Type I models). 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., publications, external funding). 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 academically useful research into tactics for cultivating citations,” for example (von Tunzelmann & Mbula, 2003, p. 15).

As illustrated by the national systems presented in this chapter, research evaluation as conducted throughout the world can be characterized by increasing levels of size and complexity. However, most countries still regard their systems as experimental. There is a near worldwide interest in the United Kingdom model, which has become a “benchmark for research evaluation” (von Tunzelmann & Mbula, 2003, p. 6). Conversely, there has been some suggestion that the United Kingdom’s RAE does not itself lead to enhancements in the quality of research in the United Kingdom, but does encourage universities and departments to compete with one another, for example, “by [universities and departments] bidding to attract star researchers in order to improve their record of achievement” (Barker & Lloyd, 1997, p. 56).
CHAPTER III

METHODOLOGY FOR COMPARATIVE EVALUATION OF INTERNATIONAL RESEARCH EVALUATION MODELS

While the previous chapter described the historical background, the research context, and the methods employed in 16 countries to evaluate publicly-funded research, this chapter enumerates the methodological approach used to assess the merits of those models/systems. The specific aims and objectives of the study are to:

1. Identify a small set of relevant and demonstrable properties that are adequate to characterize a good research evaluation model for large-scale evaluations of publicly-funded research

2. Characterize each of the countries' research evaluation models in terms of these properties that are adequate to characterize a good research evaluation model for large-scale evaluations of publicly-funded research

3. Determine the relative and absolute merits of the countries' research evaluation models for large-scale evaluation of publicly-funded research through a scoring, profiling, and synthesis procedure

4. Draw valid conclusions about how governments should evaluate publicly-funded research

The second and third of these aims and objectives are descriptive (i.e., what is), whereas the first and fourth are prescriptive (i.e., what ought to be). Specifically, the intent of the study was to assess the merits of the model from each country, to evaluate their strengths and weaknesses, to determine how they can be improved, and to ascertain what can be learned from the better models. In part, the underlying rationale of the study
was that “identification of the problem is a long step toward a solution.”

Method

The methods used to accomplish the aforementioned aims and objectives are set forth in the remainder of this chapter, these include: the design of the study; the selection and characteristics of judges who evaluated the 16 national-level research evaluation models; the study’s setting and materials; its measures; the reliability and validity of those measures; the timing and sequence of the procedural aspects of the study; and the analytic approaches applied to determine the relative and absolute merits of the national-level research evaluation models.

Study Design

The design of this study was developed on the basis of various comparative research and evaluation methodologies (Davidson, 2001; Mrinalini & Nath, 2006; Przeworski, 1987; Przeworski & Teunc, 1970; Scriven, 1991, 2005c, 2006g; Vartiainen, 2002; Weiss, 1972). Comparative evaluation refers to research in which an evaluation and the findings of the evaluation are set in a comparative framework (Vartiainen, 2002). Use of the comparative method can be justified in several ways, for example, as a means of analyzing similarities and differences in systems, in order to comprehend, to explain, or to interpret different phenomena or systems (Salminen & Lehtinen, 1982). However, the methodological principles on which the comparative evaluation process should be based are not always clear. There are, in any case, four basic principles common to the method of comparative evaluation, despite this lack of methodological clarity: (1) selection of the evaluation object; (2) the level of comparison; (3) clarification of

292 See also the Comparative Evaluation subsection of Chapter I.
characteristics used for comparison; and (4) a logical analytic framework. Each of these
design principles is briefly outlined below, and described in greater detail in other
sections of the chapter.

In this study, the evaluation objects (i.e., country research evaluation models)
were selected from the known sampling frame of two-hundred seventy-two nations,
dependent areas, and other entities in the world (Central Intelligence Agency, 2006) on
the basis of two criteria: (1) adequacy of information regarding the funding and evaluation
of research and (2) availability of that information in the English language. This sample
corresponds to more than two-thirds of the world’s top purchasing powers (i.e., GDP),
as well as a large majority of the world’s ‘research superpowers’ in terms of their
scientific productivity and government monies dedicated to research and R&D (DEST,
2003; European Commission, 2003; Group of Eight, 2002; Ministry of Research,
Ranking of World Countries, 2006). Moreover, these 16 nations produced nearly 80% of
the world’s total scientific publications during the period of 1998-2002.

The units, or level, of comparison were national-level research evaluation models
in which the similarity or dissimilarity of the systems varied widely. This similarity or
difference of the cases is just as important in comparative evaluation as it is in other
comparative studies where the assumption is that when similar cases are studied, the
differences rather than similarities are to be analyzed. The same applies when
comparative evaluation focuses on features common to cases that are different.
Comparative studies of a practical nature, however, are rarely situated at the extremes of
the continuum where one is often faced with the undesirable task of comparing ‘apples
and oranges.’ This is because many cases that appear similar involve many differing
characteristics. Every object selected for study has its own history and identity and this
poses serious difficulties when the cases are classified on the basis of similarity (Ragin, 1987).

For a comparative evaluation to be relevant, it is crucial that the comparative concepts are defined clearly so that different right-to-know audiences (e.g., researchers, policy makers) will interpret and understand the concepts involved in the same way; that is, conceptual comprehensibility. How the operation of indicators and criteria are defined plays a central role in interpreting the findings of a comparative evaluation. Definition of the concepts applied in the study is important for the comparative evaluation, since comparison makes the analysis more demanding. The analysis becomes increasingly demanding particularly because, in a study based on a comparative evaluation, the concepts are utilized not only during evaluation and analysis, but also when reporting the results of the analysis. The indicators and criteria applied in this study are presented in the Measures subsection of this chapter.

In comparative evaluation, analysis of the findings depends on the method, but it also depends on the level of the evaluation. Comparative evaluation produces comparable information most efficiently when the units analyzed are as similar as possible. This assumption is based on the idea that it is easier to form reliable evaluation criteria when evaluating similar units. In other words, when evaluating systems or organizations that are structurally, functionally, or culturally very different, one often has to operate with excessively nebulous concepts and criteria. It has been suggested that one solution to the heterogeneity problem is to determine how indicators and criteria can be weighted in a way that adequately reflects their importance (Vartiainen, 2002). Yet, without adequate justification or overwhelming evidence, equal weighting of indicators and/or merit-defining criteria should not be abandoned. The central issue, however, is not weighting, but whether or not one has identified the right indicators and the right
criteria. These issues are given greater attention in the Measures and Analytic Approach subsections of this chapter.

In more traditional research design terminology, even if comparative, this study was principally quantitative in that the qualitative properties or attributes of the merits of each country's research evaluation system were established through a measurement process in which those qualitative properties were expressed numerically (Bickman & Rog, 1998). It was also a nonexperimental study. The clearest way to classify nonexperimental quantitative research is on the basis of two key dimensions (Johnson, 2001). The first is the major or primary research objective. The research objective normally consists of three categories: descriptive (i.e., what is); predictive (i.e., what will be); and explanatory (i.e., the reason for or cause of). A better division of this classification system, however, would include the category 'evaluative.' That is, the objective of the evaluative category would be research that seeks to determine "so what?" or "what ought to be."293, 294

On this new four-part dimension, then, the present study is best classified as both descriptive and evaluative research in that the objective was to give an account of "what is" as well as to determine "what ought to be." The second, and equally useful, dimension for classifying research is time. The time dimension normally has three categories: retrospective (i.e., backward looking); cross-sectional (i.e., a single point in time); and longitudinal (i.e. more than one point in time or across time). On the time dimension, this study is best classified as retrospective. In retrospective research,

293 The problem with the current classification system is that "what is" is often contrasted with "what ought to be"; that is, a normative distinction. However, the descriptive dimension of "what is" can be merely descriptive or it can be evaluative in that it seeks answers to the questions "so what?" or "what ought to be?" or, sometimes it seeks both.

294 This new category would also include the objectives "now what?" "how much?" and "what if?" as much of evaluation involves questions related to apportioning (see Chapter I). Collectively, questions of this type are referred to as prospective, and they can be distinguished from questions about what is happening now or what has happened in the past; that is, retrospective questions.
comparisons are made between the past—as estimated by the data—and the present for the cases in the data set. Thus, this was a retrospective, descriptive, and evaluative study.

 Judges

Judges were responsible for independently and collectively rating the research evaluation models of the 16 countries. Two groups of judges participated in this study. The two groups consisted of six judges from Western Michigan University’s (WMU) Interdisciplinary Ph.D. in Evaluation (IDPE) program in Kalamazoo, Michigan, United States (US) and six from The University of Auckland (UA) in Auckland, New Zealand (NZ), for a total of twelve judges. The selection of judges who participated in the study was non-random. Rather, it was both a convenience and purposive sample. It was a sample of convenience in that judges were selected on the basis of availability and accessibility. It was purposive, however, in that the sampling procedure specifically sought to select judges on the basis of certain characteristics; that is, judges demonstrating a degree of competence in terms of—either or both—evaluation or research knowledge, skill, and ability, to allow increased diversity. Therefore, it was also a criterion-referenced sample in terms of the judges who participated (Patton, 1990) in that each potential judge was required to submit their curriculum vita (CV) or résumé for review prior to participating in the study. Prospective judge’s CVs and résumés were used to gauge their relevant professional experience, educational background, and disciplinary expertise. Judge’s CVs and résumés were also used as a basis for assigning judges to small multidisciplinary subpanels in the second stage of the study.

295 See the Stage I and Stage II subsections of this chapter.
296 Additional characteristics used to assess judge competency are presented in the Setting and Materials subsection of this chapter.
297 See the Procedure subsection of this chapter.
Combined, the average age of US \((N = 6)\) and NZ \((N = 6)\) judges was 41.41 years \((SD = 7.80)\). Of these, 67% were male and 33% were female. The majority of judges were Caucasian/White (83%); however, a small minority was Asian (8%) and Māori (8%). Nearly a dozen countries of origin were represented by the panels’ judges, including for example, Canada, Iran, Korea, New Zealand, Scotland, and the United States. Overall, 58% of judges held a Master’s degree and 42% held PhDs; also, all US judges were enrolled in a doctoral program in evaluation.

Collectively, these judges reported having a great deal of evaluation \((M = 3.58, SD = 0.80)\) and research experience \((M = 3.08, SD = 0.78)\).\(^{298}\) Moreover, their cognate or disciplinary areas of expertise and interest were widely varied and included public health, nuclear technology, education, epidemiology, psychometrics, human resources, business and industry, marketing, and econometrics, for example. In addition, these judges have published more than 30 peer reviewed papers and led or participated in equal numbers of research studies and evaluations. Table 18 presents the judges’ basic socio-demographic characteristics by US and NZ panels.

\(^{298}\) From 1 to 5, where 1 = none and 5 = extensive (see Appendix B).
Table 18
Socio-Demographic Characteristics of Judges by Panel

<table>
<thead>
<tr>
<th></th>
<th>US panel</th>
<th>NZ panel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong>*</td>
<td>37.16 (7.75)</td>
<td>45.67 (7.84)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>67%</td>
<td>67%</td>
</tr>
<tr>
<td>Female</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td><strong>Race/ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American/Black</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Asian</td>
<td>17%</td>
<td>0%</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Caucasian/White</td>
<td>83%</td>
<td>83%</td>
</tr>
<tr>
<td>East Asian/Middle Eastern</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Māori</td>
<td>0%</td>
<td>17%</td>
</tr>
<tr>
<td>Native Hawaiian</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Pasifika</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Highest academic degree held</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Master's degree</td>
<td>100%</td>
<td>17%</td>
</tr>
<tr>
<td>Doctoral degree</td>
<td>0%</td>
<td>83%</td>
</tr>
<tr>
<td><strong>Evaluation experience</strong></td>
<td>3.33 (0.82)</td>
<td>2.83 (0.75)</td>
</tr>
<tr>
<td><strong>Research experience</strong>*</td>
<td>3.17 (0.98)</td>
<td>4.00 (0.63)</td>
</tr>
</tbody>
</table>

* The standard deviation is given in parentheses.

** From 1-5, where 1 = none and 5 = extensive (see Appendix B). The standard deviation is given in parentheses.

*** From 1-5, where 1 = none and 5 = extensive (see Appendix B). The standard deviation is given in parentheses.
Although judges started by working independently, they later worked in panels and subpanels. Panels were simply the totality of judges at a given site (US and NZ). Thus, there were two panels consisting of six judges. Subpanels were simply divisions within panels (e.g., small groups of judges working as teams). In this study, pairs of judges, subpanels, and panels were treated as dyads. That is, the judges, subpanels, and panels were assumed to be nonindependent (Kenny, Kashy, & Cook, 2006). Nonindependence simply means that they (judges, subpanels, and panels) share something in common; that is, that they made judgments on the same country or countries. Each judge, subpanel, and panel was linked to one, and only one, other judge, subpanel, or panel. This linkage is one in which the judge, subpanel, and panel dyads never interacted at all and were not even aware of each other, but they were both exposed to the same stimuli.

Setting and Materials

The study was conducted at WMU’s The Evaluation Center (EC) in Kalamazoo, Michigan, United States and the UA’s School of Education in Auckland, New Zealand in order to expose and/or reduce any monocultural biases. The EC is an internationally recognized research and development center, whose mission is to advance the theory, practice, and utilization of evaluation. The EC’s principal activities are research, development, dissemination, service, instruction, and national and international leadership in evaluation. The EC, which is home to the Joint Committee on Standards for Educational Evaluation, was originally founded at The Ohio State University in 1963 and moved to WMU in 1973. The UA’s Faculty of Education is rated the top research department in the New Zealand PBRF ratings (Tertiary Education Commission, 2003b). At that time it consisted of about two-hundred thirty academics, teaching undergraduate to doctoral programs (about one-hundred forty doctoral students), had over $20 million
NZD ($14 million USD) per year in research grants, and was the home of the invention of Reading Recovery (Clay, 1993). It has recently amalgamated with a local teachers’ college.

A variety of materials were used in conducting the study, including: (1) blinded country narratives; (2) a judge socio-demographic questionnaire; (3) scoring sheets; (4) scripts; and (5) WMU Human Subjects Institutional Review Board (HSIRB) approved consent forms (Project Number 07-01-05).

The blinded country narratives were derived from the descriptions presented in Chapter II, with all pertinent identifiers removed in an effort to reduce various types of bias (Birnbaum & Stegner, 1979). Moreover, a substantial proportion of the narratives were independently verified for their accuracy by local experts. It is also assumed that judges had no serious conflicts of interest—another source of bias, but not fatal to objectivity—in terms of rating the countries. Blinding of the narratives is understood as having eliminated, or at least greatly reduced, this possibility. Judges were not asked to disclose this information (i.e., conflicts of interest). These narratives were assigned an alphanumeric code (e.g., NZ09) and matching alphanumeric codes were also applied to the ‘research landscape’ (see Table 1, p. 113) in order to provide greater contextual detail (e.g., the country’s research budget, number of researchers, and so forth).

The country models’ ‘primary purposes,’ ‘basic units of assessment,’ ‘core methods,’ and ‘key indicators’ (see Chapter II) were also alphanumerically coded and included with the blinded country narratives. Even though the narratives were blinded, they provided adequately detailed descriptions of the research context and the methods used to evaluate publicly-funded research (both ex ante and ex post) for each of the 16

299 A comparative evaluation should take into account the evaluation object’s social environment and its structures and systems. In practice, this means that the comparative evaluation process should pay attention to factors such as resources, implementation, results, social norms, and the like (Vartiainen, 2002).
countries. On average, the narratives were 15-20 pages in length for any given country.

The socio-demographic questionnaire (see Appendix B) was used to gather information about the characteristics of the judges who participated in the study, including for example, their age, their race/ethnicity, their country of origin, their highest educational degree held, and the extent and area of their professional research and/or evaluation experience. These data were intended as a supplement to the information obtained from judge’s CVs and résumés. The socio-demographic questionnaire consisted of nine open- and close-ended items and was administered to judges prior to the study by means of a Web-based survey system. All judges completed the questionnaire.

The scoring sheet (see Appendix C) consisted of 25 items (i.e., indicators) which were to be scored/rated by judges from 0-10, as well as one additional item in which judges were asked to rate the model overall (i.e., their “overall best judgment”) and one item which asked judges to ‘guess’ which country they had rated in order to investigate whether the blinding of the country narratives had worked effectively. The judge scoring sheet also included a ‘narrative critique’ section consisting of four open-ended items, which were designed to identify:

1. The most salient features and the primary reasons for judges assignment of scores/ratings to the model in question
2. Judge’s assessment of the key strengths and weaknesses of the model in question
3. Judge’s perceptions of ‘what was missing’ from the model in question
4. Judge’s suggestions or recommendations for improving the model in question

300 These blinded narratives are not included in this dissertation in order to conserve space. However, they did not differ dramatically from the descriptions presented in Chapter II except that key identifiers were removed.
301 See the *Measures* subsection of this chapter.
Scripts were developed for both stages of the study. These scripts were used in order to maintain consistency (i.e., standardization of conditions) in the Stage I and Stage II procedures at both the US and NZ sites.\textsuperscript{302} Scripting is fairly common in experimental studies and is used to reduce the likelihood of accidentally treating groups differently. Moreover, scripting increases internal control and decreases the likelihood of introducing confounds. These scripts contained all of the information about what the experimenter was to say or do during the two stages of the study, beginning with the greeting of the judges and ending with the debriefing.

Prior to participating in the study, submitting CVs or résumés, or completing the socio-demographic questionnaire, judges were provided a complete description of the study, including its intent, as well as an overview of the tasks that they would be asked to complete. This description of the study also included an HSIRB consent form. These HSIRB consent forms (see Appendices D and E) were completed by all judges prior to participation in the study.\textsuperscript{303} All materials used were original and developed specifically for the study.

Measures

Good evaluations should be \textit{valid, credible, useful, cost-effective,} and \textit{ethical}. In other words, evaluation should be logically correct and produce justifiable conclusions, be believable or have reasonable grounds for being believable to relevant audiences, be useful or designed for use, be economical in terms of the benefits produced by it, and be conducted in an ethical, legal, professional, and otherwise appropriate manner. The requirement of comprehensibility/clarity is presupposed by several of the criteria, hence

\textsuperscript{302} See the \textit{Procedure} subsection of this chapter.

\textsuperscript{303} HSIRB approval for this study can be found in Appendix F. Consent forms used for the US judges and the NZ judges can be found in Appendices D and E, respectively.
These five elements, referred to here as metadimensions (i.e., merit-defining criteria for a good research evaluation model), can be conceptually viewed as latent constructs or variables (Kelloway, 1998; Kline, 2005). Each of the five metadimensions consisted of five indicators, which are the measured or observed aspects of the metadimensions; that is, attributes or inherent qualities of the metadimensions.\textsuperscript{304}

The five metadimensions were selected through a process of logical inference rather than applying existing metaevaluation checklists (e.g., Scriven, 2005c, 2006g; Stufflebeam, 1999a, 1999b; Stufflebeam, Goodyear, Marquart, & Johnson, 2005), standards (e.g., GAO, 2003; Joint Committee on Standards for Educational Evaluation, 1988, 1994), or principles (e.g., American Evaluation Association, 2004, 2006b; Australasian Evaluation Society, 2006) on the basis that these (checklists, standards, and principles) were deemed inappropriate for the aims and objectives of the present study on the premise that they are often at such a level of abstraction so as to make them virtually useless when working at a more concrete level of analysis.

Each of the 25 indicators was rated (i.e., scored) by judges and panels from 0-10, where 0 = absence of merit and 10 = excellent.\textsuperscript{305} In the following, each of the five metadimensions and the indicators intended to measure them is operationalized.

Validity

The validity dimension, or criterion, asserts that good evaluation is logically correct and that it estimates what it is intended to estimate. Indicators applied to this dimension were:

i. Takes appropriate factors into account

\textsuperscript{304} See the \textit{Criteria versus Indicators} subsection of Chapter I.
\textsuperscript{305} See also the \textit{Scoring} subsection of this chapter.
ii. Weights factors in a transparent, defensible, appropriate manner
iii. Does not confuse grading, ranking, scoring, and apportioning
iv. Conclusions are logically and demonstrably correct, justifiable
v. Is capable of replication or has the quality of being repeatable

Credibility

The credibility dimension, or criterion, asserts that good evaluation has reasonable grounds for being believed. Indicators applied to this dimension were:

i. Is transparent
ii. Is impartial and unbiased
iii. Has defensible accounts for lack of conflict of interest
iv. Is conducted by those with adequate expertise
v. Is externally credible, believable to right-to-know audiences

Utility

The utility dimension, or criterion, asserts that good evaluation is useful or designed for use. Indicators applied to this dimension were:

i. Is relevant, fit for purpose
ii. Is timely
iii. Is easy to apply
iv. Is easy to understand
v. Provides feedback to those evaluated
Cost-Effectiveness

The cost-effectiveness dimension, or criterion, asserts that good evaluation is economical in terms of the benefits produced by it. Indicators applied to this dimension were:

i. Money costs are reasonable given the benefits produced
ii. Costs are reasonable in terms of time
iii. Costs are reasonable in terms of specialist expertise
iv. Costs are reasonable for those submitting information
v. Payoff is substantial

Ethicality

The ethicality dimension asserts that good evaluation conforms to accepted standards of conduct and that it is governed by the prima facie value of equal rights. Indicators applied to this dimension were:

i. Able to detect fraud or misconduct
ii. Allows an appeal process
iii. Deals fairly with new as well as experienced researchers
iv. Gives a complete and fair assessment
v. Is independently metaevaluated

Although there are certainly other aspects or features (i.e., indicators) of validity, credibility, utility, cost-effectiveness, and ethicality that could have been included, the aforementioned were selected on the basis that they were reasonably reliable, comprehensive, and valid for the intended purpose.
Reliability and Validity of Measures

For this study it was necessary to establish three types of reliability. At the very least, scoring had to be reliable (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This means simply that the errors of measurement were minimized by judges being systematic in their ratings and rating on a clearly applicable rubric. Given that the procedure involved ratings by others it was also necessary to establish interrater reliability and internal consistency.

Interrater, or, in this case interjudge and interpanel, reliability establishes the agreement between raters using the same measure (i.e., indicators) on the same objects (i.e., countries). These reliability estimates represent judge and panel generalizability for ratings, and were estimated by Pearson's product-moment correlation coefficients and intraclass correlation coefficients between the ratings of pairs of judges and panels; that is, dyads (Kenny, Kashy, & Cook, 2006). Thus, for each country, the judge pair consisted of one NZ and one US judge who rated the same set of countries. For panels, however, there was only one pair (i.e., NZ and US), and this pair rated all of the countries. These estimates are presented in Chapter IV.

Cronbach's coefficient $\alpha$ (Cronbach, 1951), which is often referred to as internal consistency, was used to assess the minimal standard error of measurement. Cronbach's coefficient $\alpha$ provides an index of the mean interitem correlation across judges, and this reflects the lower-bound of the estimate of reliability for the scores. These estimates are

---

306 See the Procedure subsection of this chapter.
307 US and NZ judges were randomly assigned countries to rate (see the Procedure subsection of this chapter). However, given that judges were to rate either two or three countries, both panels had the same cluster of countries assigned to their judges. For example, if US judge A (or B, or C) was assigned Belgium, Germany, and Japan, then NZ judge Z (or Y, or X) was also assigned Belgium, Germany, and Japan in order to maintain dyad-pairs so that interjudge and interpanel reliability estimates could be produced.
308 See the Procedures subsection of this chapter.
presented in Chapter IV.

Validity, on the other hand, is "...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (Messick, 1989, p. 13). Thus, the validity of this assessment is a function of the utility of the consequences, built on the adequacy of the scores (Messick, 1996a, 1996b). And, there are four parts: (1) the evidential basis of interpretation from construct validity; (2) the evidential basis of use from construct validity incorporating relevance and utility; (3) the consequential basis of interpretation from construct validity and value implications; and (4) the consequential basis of use from construct validity, relevance, utility, and the value and social consequences. These four are related to the adequacy of the criteria (see Measures), materials (see Materials), and training (see Procedure).

Procedure

Conceptually, the study procedure was developed along the fundamental principles and procedures set forth in the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996, June; Lewis, Green, Mitzel, Baum, & Patz, 1998, April) and Delphi techniques (Gordon & Helmer, 1964; Linstone & Turnoff, 2002). That is, it was aimed at building consensus through an explicit, systematic process of deliberation. However, these principles and procedures were modified to suit the specific aims of this study. This modification included a two-stage design, among others. The first stage was aimed at the calibration of judges through a training and independent rating procedure and the second set forth to achieve agreed-upon group (i.e., panel) ratings through a deliberative consensus-seeking procedure (Davis, 1996; House & Howe, 2000a, 2000b). Although the study sought comparability between the US and NZ panels (i.e., procedural
consistency), it was also responsive (Greene & Abma, 2002; Stake, 1980, 2004) in that minor improvements and refinements were made following a beta-test (i.e., pilot) of the materials and the sequence and timing of procedures at the US site (with a different set of judges) prior to full implementation.

Both stages of the study were completed in one day by both panels. The first stage began at 8:30 AM and ended at 12:30 PM. The second stage began at 2:30 PM and ended at 6:30 PM. A 15-minute break was given during each stage and a 2-hour lunch break was given between the first and second stage of the study. The US panel study took place on March 31, 2007 and the NZ panel study took place on April 12, 2007. Outlined below is the timing and sequence of the procedures which were applied to both the US and NZ panels.

Stage I

The first stage of the study involved two phases: (1) training and calibration of judges and (2) judges independently rating randomly assigned countries. The general procedure and chronological order for the two phases of Stage I was:

Phase 1
1. Introduction to the nature and intent of the study
2. Instructions for using the scoring sheet
3. Working through a hypothetical case
4. Discussion of the hypothetical case
5. Instructions for completing independent ratings

Phase 2
6. Random assignment of countries to judges
7. Independent ratings by judges
In all, Stage I was 4 hours in length and consisted of the seven key tasks listed above. First, judges were introduced to the general nature and intent of the study. Second, judges were given detailed instructions for using the scoring sheets. Third, judges were randomly divided into three groups of two (i.e., subpanels). These subpanels then worked through a hypothetical case using the scoring sheet because agreement frequently drops when judging 'real' data (Romanczyk, Kent, Diament, & O'Leary, 1973). Fourth, an interactive discussion was held to discuss the hypothetical case. The discussion focused on discrepancies in ratings to highlight the reasons implicit in each judge's and subpanel's ratings and to work toward a shared set of parameters underlying the ratings; that is, this task was designed so that all judges are calibrated in the same way and so that a numerical rating will represent the same cognitive appraisal by different judges. Lastly, judges were given a final set of instructions for completing their independent ratings.

The second phase of Stage I, immediately following the general training and calibration procedures, involved randomly assigning the complete set of 16 countries to judges. Four of the panel judges were randomly assigned three countries to rate and two judges were randomly assigned two. Judges were provided blinded country narratives and scoring sheets for each of their randomly assigned countries. Judges were instructed to work independently and were given the remainder of the stage to complete the independent rating task.

At all times throughout the session, judges were encouraged to ask questions. During the 2-hour lunch break, all Stage I data were entered and analyzed in preparation

---

309 This could potentially produce what is known as a practice effect (Shepard, 1993), in that some judges received two countries to rate and others received three. The usual assumption underlying practice effects are that subjects (i.e., judges) 'learn how to do the task and subsequent performance improves.' However, the design of the study explicitly sought improved performance, not to avoid it. That is, the underlying methodological assumptions were that the study specifically sought panel consensus.

310 See the Setting and Materials subsection of this chapter.
Stage II

The procedure for Stage II was:

1. Overview of Stage I results
2. Subpanel formation
3. Re-rating by subpanels
4. Discussion of new ratings
5. Deliberation and consensus

Stage II involved five procedures, and like Stage I was 4 hours in length. First, an overview was given of the Stage I ratings. Second, three subpanels consisting of two judges were formed. These subpanels were formed *a priori* on the basis of disciplinary expertise and evaluation and/or research experience as determined through the socio-demographic questionnaire and judge’s CVs and résumés. The judges brought their completed ratings of their two or three randomly assigned countries to the subpanel. Third, the newly-formed subpanel re-rated judges’ randomly assigned countries using the Stage I results as a starting point. These groups of judges worked as teams by country, and made a decision to resolve any discrepancies from Stage I. They then agreed to a small set of reasons for putting a country into a rating category. Each group presented their ratings and the ratings were discussed by the entire panel. Finally, the entire panel deliberated the subpanel ratings and decided on a final rating on each indicator for each country, which was intended to reflect the collective decision of all judges within the panel. As in Stage I, judges, subpanels, and panels were encouraged to ask questions at all time throughout the session.

---

311 See the *Judges* subsection of this chapter.
Figure 6 illustrates the design of the study’s procedure. In this figure, the solid lines indicate procedures within judges, subpanels, and panels. Dashed lines indicate procedures between judges, subpanels, and panels.

![Diagram](image)

Figure 6. Procedural Design of Comparative Evaluation

**Analytic Approach**

Four analytic approaches were applied to judges’ and panel’s ratings: scoring; profiling; synthesis; and human judgment analysis. Scoring was used to amalgamate judges’ and panel’s indicator ratings to a score on each of the five dimensions. Profiling was used to develop a multidimensional profile of each country in terms of their validity, credibility, utility, cost-effectiveness, and ethicality. A synthesis procedure was used to convert the metadimensional performances of each country to an overall evaluative conclusion about each country’s research evaluation model. Finally, human judgment analysis was used to investigate how each judge weighted model’s attributes.
Scoring

Each of the five indicators on each of the five metadimensions was rated (i.e., scored) by judges and panels from 0-10, where 0 = absence of merit and 10 = excellent. Each of the five metadimensions was weighted equally at 20% of the total 'merit space' as shown in Table 19.\textsuperscript{312} The sum of the weights equals 100%. This table also illustrates the basic procedure by which raw scores and weighted scores were calculated for each of the five metadimensions.

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>Indicator</th>
<th>Raw score</th>
<th>Weight</th>
<th>Weighted score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Validity</td>
<td>i</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iv</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>v</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sum (i, ii, iii, iv, v) = 0.50$</td>
<td>20% (2)</td>
<td>0-100</td>
<td></td>
</tr>
<tr>
<td>2. Credibility</td>
<td>i</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iv</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>v</td>
<td>0-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sum (i, ii, iii, iv, v) = 0.50$</td>
<td>20% (2)</td>
<td>0-100</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{312} All of the metadimensions were given equal weighting as there was no reasonable or demonstrable justification for claiming that validity is more important than cost-effectiveness, or that ethicality is more important than credibility, and so forth.
Table 19—Continued

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>Indicator</th>
<th>Raw score</th>
<th>Weight</th>
<th>Weighted score</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Utility</td>
<td>$\sum (i, ii, iii, iv, v) = 0-50$</td>
<td>20% (2)</td>
<td>0-100</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iv</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Cost-effectiveness</td>
<td>$\sum (i, ii, iii, iv, v) = 0-50$</td>
<td>20% (2)</td>
<td>0-100</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iv</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Ethicality</td>
<td>$\sum (i, ii, iii, iv, v) = 0-50$</td>
<td>20% (2)</td>
<td>0-100</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iii</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iv</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>0-10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus, if country X received a rating of 5 on indicator $I$, a rating of 7 on indicator $\bar{ii}$, a rating of 6 on indicator $\bar{iii}$, a rating of 5 on indicator $iv$, and a rating of 9 on indicator $v$, on the validity metadimension, for example, the raw score would be 32 and the weighted score would be 64, or 64%, as shown in Equations 5 and 6\(^{313}\)

Raw Score = $\sum(I_i, I_{\bar{ii}}, I_{\bar{iii}}, I_{iv}, I_v)$

= $(5 + 7 + 6 + 5 + 9) = 32$ \hspace{1cm} (5b)

Weighted Score = Raw Score $\times$ Weight \hspace{1cm} (6a)

\(^{313}\) Thus, each indicator accounted for 20% of the weighted score for any given metadimension.
\[ = 32 \times 2 = 64 \]  
\[ r = \frac{WS}{TPS} = \frac{32}{50} = 64\% \]

where \( I \) is indicator, \( WS \) is weighted score, and \( TPS \) is total possible score on any given metadimension (i.e., 0-50).

**Profiling**

Profiling was used to graphically exhibit *grades* (see ‘quality categories’ in the *Synthesis* subsection of this chapter) and *scores* on the relevant dimensions of merit (i.e., validity, credibility, utility, cost-effectiveness, and ethicality). A hypothetical example of a country profile showing metadimensional scores and the total weighted score is shown below in Figure 7.

![Hypothetical Country Profile](image-url)
Synthesis

A modified NWS (see Chapter I) procedure was used to synthesize metadimension scores to a total weighted score or average total weighted score. The usual NWS procedure was modified in an effort to avoid the problems typically associated with NWS procedures (Persaud, 2006, November) by weighting each of the indicators and metadimensions equally. From the five metadimension weighted scores it was possible to compute both a total weighted score and an average total weighted score as shown in Equations 7 and 8\(^{314}\)

\[
\text{Total Weighted Score} = \sum (MD_1, MD_2, MD_3, MD_4, MD_5) = 0-500 \tag{7a}
\]

or \[
\frac{TWS}{TPMDS} = 0\%-100\% \tag{7b}
\]

\[
\text{Average Total Weighted Score} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{8a}
\]

\[
= \frac{1}{5} (MD_1 + MD_2 + MD_3 + MD_4 + MD_5) \tag{8b}
\]

\[
= \frac{(MD_1 + MD_2 + MD_3 + MD_4 + MD_5)}{5} = 0-100 \tag{8c}
\]

or \[
\frac{ATWS}{TPMDS} = 0\%-100\% \tag{8d}
\]

where \(MD\) is metadimension, \(TWS\) is total weighted score, \(ATWS\) is average total weighted score, and \(TPMDS\) is total possible score across all metadimensions (i.e., 0-500).

\(^{314}\) Thus, each of the twenty-five indicators accounted for 4\% of the total weighted score and each of the five metadimensions accounted for 20\% of the total weighted score.
The rationale underlying the scoring and synthesis procedures was that discrimination between the observed merit of countries on any meta-dimension, or as a whole (i.e., ranking), would be more precise by providing greater ranges of possible merit than would be obtained using, for example, nominal or ordinal, Likert-type scales (e.g., 1 = poor, 2 = fair, 3 = good, 4 = very good, and 5 = excellent)\textsuperscript{315} or binary, dichotomous types of measurement (e.g., 0 = absence of merit, 1 = presence of merit). Nevertheless, there is a point constancy (Scriven, 1991) underlying the numerical scoring schema.\textsuperscript{316} This synthesis procedure also allowed cross-case analysis (Yin, 1994) and ranking of the country models. In aiming for greater precision (i.e., 0-10) the scoring schema sacrifices some degree of reliability (Crocker & Algina, 1986; McDonald, 1999) that may have otherwise been increased by using a narrower, more restricted response range. This was, however, a conscious and concerted decision based on the comparative nature of the study, its research questions, and the analytic procedures and approaches developed to make between-country comparisons. Moreover, constant calibration procedures applied throughout Stage I and II was specifically aimed at solidifying judges' decisions and increasing reliability.

The total weighted scores on each of the five metadimensions were then assigned to a quality category (i.e., a grade or rating), as presented in Tables 20 and 21. It is also possible to apply the quality categories to the average total weighted scores. A total weighted score on any of the five meta-dimensions $\leq 50$, or $\leq 50\%$, constitutes failure on the dimension, as does an average total weighted score of $\leq 50$, or $\leq 50\%$; that is, a bar has been placed on scores below this point. Also, a stepping procedure (see Chapter I) was applied to converting weighted scores to quality categories in that the upper

\textsuperscript{315} Likert-type scales are not useful for most mathematical operations (e.g., sum, mean) as the distance between categories is usually unknown and/or unequal.

\textsuperscript{316} See Chapter I for a more detailed discussion of the point constancy requirement.
quality categories are incrementally smaller (and more difficult to achieve) than lower quality performances. For example, the A, B, and C categories have ranges of 10, whereas the D category has a range of 20 and the F category a range of 50.\textsuperscript{317}

Table 20

Conversion of Weighted Scores to Quality Categories

<table>
<thead>
<tr>
<th>Weighted score</th>
<th>Quality category</th>
</tr>
</thead>
<tbody>
<tr>
<td>91-100</td>
<td>A</td>
</tr>
<tr>
<td>81-90</td>
<td>B</td>
</tr>
<tr>
<td>71-80</td>
<td>C</td>
</tr>
<tr>
<td>51-70</td>
<td>D</td>
</tr>
<tr>
<td>≤ 50</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 21

Quality Category Descriptions

<table>
<thead>
<tr>
<th>Quality category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Excellent; clear example of exemplary performance; no deficiencies</td>
</tr>
<tr>
<td>B</td>
<td>Very good; strong overall but not exemplary; no real deficiencies of consequence</td>
</tr>
<tr>
<td>C</td>
<td>Good; reasonably good; minor but nonfatal deficiencies</td>
</tr>
<tr>
<td>D</td>
<td>Satisfactory; barely adequate; some serious deficiencies</td>
</tr>
<tr>
<td>F</td>
<td>Absence of merit; clearly inadequate; fatal deficiencies</td>
</tr>
</tbody>
</table>

\textsuperscript{317} See Chapter I for a more detailed discussion of the stepping procedure.
Human Judgment Analysis

Secondary analyses were conducted using elements of human judgment analysis, known as the Lens Model (Brunswick, 1943; Hammond, 1955). The Lens Model represents the relationship between a perceiver (i.e., judge) and the objects of perception, or judgment, as mediated by cues whose relationship to both the judge and the object is probabilistic. Thus an object is not itself seen; it is seen only through a set of cues. Although equal weights were used for twenty-five indicators and the five metadimensions, an ordinary least squares (OLS) multiple regression analysis was used to investigate how each judge weighted these attributes in their assessment by regressing judge’s five metadimensional raw scores on their “overall best judgment” score, or rating, in the form of the linear equation given in Equation 9

$$\hat{Y}_i = a + b_1VALIDITY_i + b_2CREDIBILITY_i + b_3UTILITY_i + b_4COST-EFFECTIVENESS_i + b_5ETHICALITY_i$$

(9)

where $\hat{Y}$ is the predicted value for a judge’s “overall best judgment” $i$, expressed in terms of the constant $a$ of the intercept term, and where $VALIDITY_i$ is the raw score of the five validity metadimension indicators for given judge $i$, where $CREDIBILITY_i$ is the raw score of the five credibility metadimension indicators for given judge $i$, where $UTILITY_i$ is the raw score of the five utility metadimension indicators for given judge $i$, where $COST-EFFECTIVENESS_i$ is the raw score of the five cost-effectiveness metadimension indicators for given judge $i$, and where $ETHICALITY_i$ is the raw score of the five ethicality metadimension indicators for given judge $i$. In this equation, the regression coefficients ($bs$), also referred to as the partial regression coefficients (Myers, Gamst, & Guarino, 2006), represent the independent contributions of each independent variable to the prediction of the dependent variable $\hat{Y}$. The results of these analyses are presented in Chapter IV.
Summary

The study described in this chapter sets forth to: (1) identify a small set of relevant and demonstrable properties that are adequate to characterize a good research evaluation model of large-scale evaluations of publicly-funded research; (2) characterize the countries' models in terms of those properties; (3) determine the relative and absolute merits of the countries' models using a process of scoring, profiling, and synthesis; and (4) draw valid conclusions about how governments should evaluate publicly-funded research. Moreover, the study aimed to: (a) assess the country model's merits; (b) critique their strengths and weaknesses; (c) determine how they can be improved; and (d) ascertain what can be learned from the better models.
CHAPTER IV

STAGE I AND II RESULTS: INDEPENDENT AND CONSENSUS RATINGS, SCORES, AND RANKINGS

This chapter presents the results of the Stage I independent rating procedure and Stage II consensus rating procedure for both the NZ and US panels in terms of judges’ and panels’ ratings of the national models, the conversion of ratings to scores, and the rankings of the national models on each of the five metadimensions as well as overall, with an emphasis on the Stage II consensus ratings. This chapter is divided into three parts: (1) primary analyses and results (e.g., judge and panel ratings, scores, and rankings); (2) secondary analyses and results (e.g., effectiveness of bias reduction via blinding, various estimates of reliability, attributes predicting judges’ ‘overall best judgment,’ the generalizability of judges’ ratings, and the key strengths and weaknesses of the national models as described by the judges); and (3) a summary of the key findings from the study.

Independent Judges’ and Panels’ Consensus Ratings

In order to determine whether ratings changed dramatically from Stage I to Stage II, paired-samples t-tests were conducted to assess the degree to which total weighted scores increased or decreased for both panels. For the NZ panel, total weighted scores

318 Preliminary results of this study were presented to the Tertiary Education Commission in Wellington, New Zealand (Coryn, 2007, April) and at Western Michigan University's Graduate College Research Day (Coryn & Scriven, 2007, April).
dropped an average of -1.93% from Stage I to Stage II; \( t(15) = .52, p = .61 \). For the US panel, total weighted scores dropped an average of -2.32% from Stage I to Stage II; \( t(15) = .51, p = .62 \). Given that the magnitude of change in ratings between Stage I and Stage II were relatively small and non-significant, and that the procedure was designed to seek consensus, the results presented in this section focus primarily on the Stage II ratings, scores, and rankings.

**Primary Analyses and Results**

The primary analyses and results presented here focus on judge and panel ratings, scores, and rankings of the 16 national models overall, as well as on each of the five metadimensions and concludes with assignment of the models to quality categories.

**Validity Ratings, Scores, Rankings, and Profiles**

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 \( r_1 = .90 \) (\( df = 30, p < .01 \)). The country-by-country validity weighted scores for both panels are shown graphically in Figure 8 and in tabular form in Table 22.

As illustrated in the figure and table, 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 8. Profile of Validity Weighted Scores

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

As suggested by the distribution of validity weighted scores, and illustrated in the figure, these scores were highly variable between countries. However, the difference between panel's judgments on the same countries was generally quite small—on average, 0.81%; \( t(15) = .34, p = .74 \). Only one country—New Zealand—crossed the minimum threshold for an “A” rating (see Tables 20 and 21) in terms of the model’s validity.

Potential reasons for the variability in validity scores—as suggested by panel judges—are presented elsewhere in this chapter and in Chapter V. Presented in the next section are the composite weighted credibility scores.
Table 22

Validity Weighted Scores

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>68</td>
<td>58</td>
<td>34</td>
<td>53</td>
<td>36</td>
<td>6</td>
<td>60</td>
<td>30</td>
<td>36</td>
<td>36</td>
<td>78</td>
<td>86</td>
<td>40</td>
<td>44</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>US</td>
<td>58</td>
<td>34</td>
<td>40</td>
<td>50</td>
<td>32</td>
<td>10</td>
<td>56</td>
<td>46</td>
<td>36</td>
<td>44</td>
<td>66</td>
<td>92</td>
<td>42</td>
<td>38</td>
<td>72</td>
<td>80</td>
</tr>
<tr>
<td>Differential</td>
<td>±10</td>
<td>±24</td>
<td>±6</td>
<td>±3</td>
<td>±4</td>
<td>±4</td>
<td>±16</td>
<td>±0</td>
<td>±8</td>
<td>±12</td>
<td>±6</td>
<td>±2</td>
<td>±6</td>
<td>±0</td>
<td>±8</td>
<td></td>
</tr>
</tbody>
</table>

Note. *The possible range of validity weighted scores was from 0-100, or 0%-100%.*
Credibility Ratings, Scores, Rankings, and Profiles

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 $r_1 = .91$ ($df = 30, p < .01$). The country-by-country credibility weighted scores for both panels are shown graphically in Figure 9 and in tabular form in Table 23.

As illustrated in the figure and table, 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 ($\pm 12$) and the United Kingdom ($\pm 24$).
As suggested by the distribution of credibility weighted scores, and illustrated in the figure, these scores were highly variable between countries. However, the difference between panel’s judgments on the same countries was generally quite small—on average, 1.18%; \( \tau(15) = .59, p = .56 \). Again, only one country—New Zealand—crossed the minimum threshold for an “A” rating (see Tables 20 and 21) in terms of the model’s credibility.

Potential reasons for the variability in credibility scores—as suggested by panel judges—are presented elsewhere in this chapter and in Chapter V. In the next section, the composite weighted utility scores are presented.
Table 23

Credibility Weighted Scores

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>58</td>
<td>36</td>
<td>38</td>
<td>50</td>
<td>34</td>
<td>12</td>
<td>60</td>
<td>34</td>
<td>44</td>
<td>42</td>
<td>80</td>
<td>88</td>
<td>36</td>
<td>51</td>
<td>64</td>
<td>70</td>
</tr>
<tr>
<td>US</td>
<td>58</td>
<td>46</td>
<td>42</td>
<td>46</td>
<td>36</td>
<td>14</td>
<td>58</td>
<td>38</td>
<td>44</td>
<td>42</td>
<td>74</td>
<td>92</td>
<td>48</td>
<td>44</td>
<td>88</td>
<td>66</td>
</tr>
<tr>
<td>Differential</td>
<td>±0</td>
<td>±10</td>
<td>±4</td>
<td>±4</td>
<td>±4</td>
<td>±0</td>
<td>±4</td>
<td>±0</td>
<td>±6</td>
<td>±4</td>
<td>±12</td>
<td>±7</td>
<td>±24</td>
<td>±4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *The possible range of credibility weighted scores was from 0-100, or 0%-100%.
Utility Ratings, Scores, Rankings, and Profiles

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%; \( \kappa(15) = .20, p = .84 \). 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 \( r_1 = .93 (df = 30, p < .01) \). The country-by-country utility weighted scores for both panels are shown graphically and in tabular form in Figure 10 and Table 24, respectively.

As illustrated in the figure and table, 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 10. Profile of Utility Weighted Scores

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

Like the validity and credibility metadimensions, and as suggested by the distribution of utility weighted scores, and illustrated in the figure, these scores were highly variable between countries. However, the difference between panel’s judgments on the same countries was generally quite small—on average, 0.37%; t(15) = .20, p = .84. As with the prior two metadimensions, only one country—New Zealand—met the minimum threshold for an “A” rating (see Tables 20 and 21) in terms of the model’s utility.

Potential reasons for the variability in utility scores—as suggested by panel judges—are presented elsewhere in this chapter and in Chapter V. Presented in the next section are the composite weighted cost-effectiveness scores.
Table 24

Utility Weighted Scores

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>36</td>
<td>42</td>
<td>44</td>
<td>40</td>
<td>34</td>
<td>6</td>
<td>64</td>
<td>30</td>
<td>46</td>
<td>46</td>
<td>66</td>
<td>86</td>
<td>56</td>
<td>46</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>US</td>
<td>54</td>
<td>42</td>
<td>44</td>
<td>50</td>
<td>34</td>
<td>0</td>
<td>54</td>
<td>38</td>
<td>44</td>
<td>38</td>
<td>70</td>
<td>90</td>
<td>48</td>
<td>46</td>
<td>74</td>
<td>66</td>
</tr>
<tr>
<td>Differential</td>
<td>±18</td>
<td>±10</td>
<td>±10</td>
<td>±6</td>
<td>±10</td>
<td>±8</td>
<td>±2</td>
<td>±8</td>
<td>±4</td>
<td>±8</td>
<td>±10</td>
<td>±2</td>
<td>±6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. *The possible range of utility weighted scores was from 0-100, or 0%-100%.
Cost-Effectiveness Ratings, Scores, Rankings, and Profiles

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 11 and in tabular form in Table 25.

As illustrated in the figure and table, 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 ($\pm 14$) and the United Kingdom ($\pm 24$).
Figure 11. Profile of Cost-Effectiveness Weighted Scores

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

As with the prior metadimensions, and as suggested by the distribution of cost-effectiveness weighted scores, and illustrated in the figure, these scores were highly variable between countries. However, the difference between panel's judgments on the same countries was generally quite small—on average, 1.37%; t(15) = .52, p = .61. Only one country—New Zealand—crossed the minimum threshold for an “A” rating (see Tables 20 and 21) in terms of the model's cost-effectiveness.

Potential reasons for the variability in cost-effectiveness scores—as suggested by panel judges—are presented elsewhere in this chapter and in Chapter V. In the next section, the composite weighted ethicality scores are presented.
Table 25

Cost-Effectiveness Weighted Scores

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>42</td>
<td>28</td>
<td>38</td>
<td>40</td>
<td>40</td>
<td>12</td>
<td>58</td>
<td>42</td>
<td>44</td>
<td>46</td>
<td>72</td>
<td>80</td>
<td>54</td>
<td>44</td>
<td>82</td>
<td>62</td>
</tr>
<tr>
<td>US</td>
<td>38</td>
<td>36</td>
<td>44</td>
<td>46</td>
<td>28</td>
<td>14</td>
<td>44</td>
<td>56</td>
<td>32</td>
<td>52</td>
<td>70</td>
<td>92</td>
<td>46</td>
<td>38</td>
<td>58</td>
<td>68</td>
</tr>
<tr>
<td>Differential</td>
<td>±4</td>
<td>±8</td>
<td>±6</td>
<td>±6</td>
<td>±12</td>
<td>±2</td>
<td>±14</td>
<td>±14</td>
<td>±12</td>
<td>±6</td>
<td>±2</td>
<td>±12</td>
<td>±8</td>
<td>±6</td>
<td>±24</td>
<td>±6</td>
</tr>
</tbody>
</table>

*Note:* The possible range of cost-effectiveness weighted scores was from 0-100, or 0%-100%.
Ethicality Ratings, Scores, Rankings, and Profiles

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%; χ(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 r₁ = .76 (df = 30, p < .01). The country-by-country ethicality weighted scores for both panels are shown graphically in Figure 12 and in tabular form in Table 26.

As illustrated in the figure and table, 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 12. Profile of Ethicality Weighted Scores

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

As suggested by the distribution of validity weighted scores, and illustrated in the figure, these scores were highly variable between countries. However, the difference between panel’s judgments on the same countries was generally quite small—on average, 2.62%; \( t(15) = .89, p = .39 \). None of the models crossed the minimum threshold for an “A” rating (see Tables 20 and 21) in terms of ethicality.

Potential reasons for the variability in ethicality scores—as suggested by panel judges—are presented elsewhere in this chapter and in Chapter V. Presented in the next section are the composite total weighted scores.
Table 26

Ethicality Weighted Scores

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>36</td>
<td>16</td>
<td>38</td>
<td>46</td>
<td>38</td>
<td>18</td>
<td>46</td>
<td>42</td>
<td>38</td>
<td>46</td>
<td>62</td>
<td>86</td>
<td>44</td>
<td>46</td>
<td>76</td>
<td>64</td>
</tr>
<tr>
<td>US</td>
<td>50</td>
<td>44</td>
<td>44</td>
<td>42</td>
<td>24</td>
<td>18</td>
<td>44</td>
<td>24</td>
<td>36</td>
<td>32</td>
<td>52</td>
<td>86</td>
<td>38</td>
<td>44</td>
<td>58</td>
<td>64</td>
</tr>
<tr>
<td>Differential</td>
<td>±14</td>
<td>±28</td>
<td>±6</td>
<td>±4</td>
<td>±14</td>
<td>±0</td>
<td>±2</td>
<td>±18</td>
<td>±12</td>
<td>±14</td>
<td>±10</td>
<td>±6</td>
<td>±6</td>
<td>±2</td>
<td>±18</td>
<td>±0</td>
</tr>
</tbody>
</table>

Note. *The possible range of ethicality weighted scores was from 0-100, or 0%-100%.*
Total Ratings, Scores, Rankings, and Profiles

The average difference between the two panels’ total weighted scores was 0.65% ($\frac{1}{15} = .68, p = .51$). 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 13 and in tabular form in Table 27. Figure 14 and Figure 15 shows the complete multidimensional profile for each of the national models in rank order (descending by total weighted scores) by panel.

Panels’ dimensional and total raw and weighted ratings were not averaged because of the small size of the sample (i.e., two panels). Therefore, consistency of the ratings between the two panels (i.e., cross-case analysis; Yin, 1994) was used to assess the final rankings of the national models. As shown in Figures 13, 14, and 15 and Table 27, the top-rated models were unanimously the Type I, large-scale performance exercises (i.e., the Netherlands, New Zealand, the United Kingdom, and the United States) of various hues and those which are generally consistent in their evaluation approach (Type A). Although Hong Kong’s RAE is both Type I and Type A, judges scored the model somewhat lower than those of New Zealand, the United Kingdom, the Netherlands, and the United States. Equally noticeable are the consistently low ratings received by the French model across all of the merit-defining metadimensions by both panels. The remainder of the models (i.e., those in the middle of the distribution) are the bulk
funding (Type II) and indicator-driven (Type III) mechanisms, which are characterized by a high degree of variability in terms of their research assessment strategies (Type B).

Figure 13. Profile of Total Weighted Scores

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

Total weighted scores, combining the five metadimension, were highly variable between countries. However, the difference between panel's judgments on the same countries was generally quite small—on average, 0.65%; \( t(15) = .68, p = .51 \). Only one country—New Zealand—managed to cross the minimum threshold for an overall "A" rating (see Tables 20 and 21).
### Table 27

**Total Weighted Scores**

<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>48.00</td>
<td>40.00</td>
<td>38.40</td>
<td>45.80</td>
<td>36.40</td>
<td>10.80</td>
<td>57.60</td>
<td>35.60</td>
<td>41.60</td>
<td>43.20</td>
<td>71.60</td>
<td>85.20</td>
<td>46.00</td>
<td>46.20</td>
<td>73.20</td>
</tr>
<tr>
<td>US</td>
<td>51.60</td>
<td>40.40</td>
<td>42.80</td>
<td>46.80</td>
<td>30.80</td>
<td>11.20</td>
<td>51.20</td>
<td>40.40</td>
<td>38.40</td>
<td>41.60</td>
<td>66.40</td>
<td>90.40</td>
<td>44.40</td>
<td>42.00</td>
<td>70.00</td>
</tr>
<tr>
<td>Differential</td>
<td>±3.60</td>
<td>±0.40</td>
<td>±4.40</td>
<td>±1.00</td>
<td>5.60</td>
<td>±0.40</td>
<td>±6.40</td>
<td>±4.80</td>
<td>±3.20</td>
<td>±1.60</td>
<td>±5.20</td>
<td>±5.20</td>
<td>±1.60</td>
<td>±4.20</td>
<td>±3.20</td>
</tr>
</tbody>
</table>

*Note.* The possible range of total weighted scores was from 0-100, or 0%-100%. 
Figure 14. Profile of NZ Panel Metadimensional and Total Ratings

Figure 15. Profile of US Panel Metadimensional and Total Ratings
Quality Categories of National Models

As shown in Table 28, and considering the overall ratings of both panels, only 25% (4 of 16) of the national models consistently met the minimum threshold for being assigned to a quality category (see Tables 20 and 21, p. 301) greater than F (i.e., absence of merit; clearly inadequate; fatal deficiencies). However, the US panel assigned Australia’s Research Quality Framework (RQF) to a quality category of D (i.e., satisfactory; barely adequate; some serious deficiencies), whereas the NZ panel rated the model an F, therefore, 31% (5 of 16) could be considered above a quality category threshold of F. Most striking in these results is the sheer number of F quality ratings (21 of the 32 total ratings, or 66%).

The nature of the large number of failures, in part, can be attributed to several key characteristics of the systems found by the panel judges to be inconsistent with a model that is valid, credible, useful, cost-effective, or ethical. In many cases, these evaluation systems are overly reliant on self-evaluation (i.e., evaluating the quality or value of one’s own work), or evaluations conducted by internal (i.e., national versus international or mixed) expert or peer review panels lacking sufficient procedures to account for conflicts of interest, serious concerns about the transparency of the process, or poorly constructed indicators of research performance (sometimes several places removed from actual research performance), for example. Most importantly, in many smaller countries as well as small or emerging scientific disciplines, is the fact that expert or peer panels often evaluate their own work or the work of colleagues because of the small number of available experts in a particular country or working in an emerging or relatively small substantive area.
<table>
<thead>
<tr>
<th>Panel</th>
<th>AU</th>
<th>BE</th>
<th>CZ</th>
<th>DE</th>
<th>FI</th>
<th>FR</th>
<th>HK</th>
<th>HU</th>
<th>IE</th>
<th>JP</th>
<th>NL</th>
<th>NZ</th>
<th>PL</th>
<th>SE</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>D</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>C</td>
<td>B</td>
<td>F</td>
<td>F</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>US</td>
<td>D</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>D</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>D</td>
<td>A</td>
<td>F</td>
<td>F</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>
Secondary Analyses and Results

As mentioned at the outset of this chapter, the secondary analyses and results presented in this section focus on the effectiveness of bias reduction via blinding, various estimates of reliability, attributes predicting judges' "overall best judgment," the generalizability of judges' ratings, and the key strengths and weaknesses of the national models as described by the judges. Information from both the Stage I and Stage II procedures were used for these analyses.

Effectiveness of Bias Reduction via Blinding

Blinding is a common method to prevent conscious and unconscious bias in many forms of research. In a single blind experiment, for instance, individual subjects do not know whether they are 'test' subjects or members of an 'experimental control' group. Single-blind experimental design is used where the experimenters either must know the complete facts or will not introduce further bias. There is a risk, however, that subjects may be influenced by interaction with the researcher, known as the experimenter's bias. Single-blind trials are especially risky in psychology and social science research, where the researcher has an expectation of what the outcome should be, or would like it to be, and may consciously or unconsciously influence the behavior of the subject (Vogt, 2005). In most cases these influences are unintentional.

Blinding was used in this study in an effort to reduce biases of various types, for example, a judge's preference (i.e., positive bias) for a particular nation or national research evaluation model, or reciprocally, negative bias toward a particular nation or national research evaluation model. For the NZ panel judges, two countries were correctly identified (one was the New Zealand model). Only one judge on the US panel
correctly identified one of the national systems rated, but not the United States’ model. In both instances, these were less than would have occurred by chance alone. Given that most judges incorrectly identified their assigned countries, positive or negative bias toward any nation or national model was negligible.

Reliability Estimates for Independent Judges and Panels

A variety of reliability estimates are presented here from both the independent rating procedure (Stage I), which was aimed at calibration of judges’ ratings, and the consensus rating procedure (Stage II), which was aimed at solidifying judges’ ratings.

The overall correlation coefficient for the correlation coefficients between NZ and US independent judges’ “overall best judgment” and each of the 25 indicators was $r = .89$ ($df = 23, p < .01$). The overall correlation coefficient for the rank orderings of the magnitude of the indicators and “overall best judgment” correlation coefficients was $r = .84$ ($df = 23, p < .01$). As shown in Table 29, all indicator to “overall best judgment” correlation coefficients were moderate to large in magnitude (i.e., from $r = .44$ to $r = .96$).

Both groups of judges gave the most weight to “Is relevant, fit for purpose” (NZ rank order = #1; US rank order = #3), “Takes appropriate factors into account” (NZ rank order = #2; US rank order = #2), and “Conclusions are logically and demonstrably correct, justifiable” (NZ rank order = #3; US rank order = #1) in their overall appraisal of the national models. Conversely, both groups of judges gave the least weight to “Deals fairly with new as well as experienced researchers” (NZ rank order = #24; US rank order = #24) and “Allows an appeals process” (NZ rank order = #25; US rank order = #25). Of these 25, only the latter (i.e., “Allows and appeals process”) correlated $< .50$ with judges’ overall judgment in their initial assessment.
### Table 29

Pearson’s Product-Moment Correlation Coefficients for Indicator Ratings and “Overall Best Judgment” and Rank Orderings of the Magnitude of the Pearson’s Product-Moment Correlation Coefficients

<table>
<thead>
<tr>
<th>Indicator</th>
<th>( r )</th>
<th>NZ</th>
<th>US</th>
<th>Differential</th>
<th>Rank</th>
<th>NZ</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is relevant, fit for purpose</td>
<td>.93</td>
<td>.94</td>
<td></td>
<td>±.01</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Takes appropriate factors into account</td>
<td>.92</td>
<td>.95</td>
<td></td>
<td>±.03</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Conclusions are logically and demonstrably correct, justifiable</td>
<td>.91</td>
<td>.96</td>
<td></td>
<td>±.05</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gives a complete and fair assessment</td>
<td>.91</td>
<td>.92</td>
<td></td>
<td>±.01</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Able to detect or uncover fraud or misconduct</td>
<td>.88</td>
<td>.88</td>
<td></td>
<td>±.00</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Money costs are reasonable given the benefits produced</td>
<td>.88</td>
<td>.88</td>
<td></td>
<td>±.00</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Weights factors in a transparent, defensible, appropriate manner</td>
<td>.88</td>
<td>.89</td>
<td></td>
<td>±.01</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Is externally credible, believable to right-to-know audiences</td>
<td>.87</td>
<td>.81</td>
<td></td>
<td>±.06</td>
<td>8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Costs are reasonable in terms of time</td>
<td>.85</td>
<td>.87</td>
<td></td>
<td>±.02</td>
<td>9</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Costs are reasonable in terms of specialist expertise</td>
<td>.85</td>
<td>.87</td>
<td></td>
<td>±.02</td>
<td>10</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Is impartial and unbiased</td>
<td>.85</td>
<td>.79</td>
<td></td>
<td>±.04</td>
<td>11</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Is conducted by those with adequate expertise</td>
<td>.84</td>
<td>.87</td>
<td></td>
<td>±.03</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Has defensible accounts for lack of conflict of interest</td>
<td>.83</td>
<td>.88</td>
<td></td>
<td>±.05</td>
<td>13</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Is timely</td>
<td>.83</td>
<td>.89</td>
<td></td>
<td>±.06</td>
<td>14</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Is capable of replication or has the quality of being repeatable</td>
<td>.82</td>
<td>.77</td>
<td></td>
<td>±.05</td>
<td>15</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Is transparent</td>
<td>.81</td>
<td>.83</td>
<td></td>
<td>±.02</td>
<td>16</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Provides feedback to those evaluated</td>
<td>.79</td>
<td>.77</td>
<td></td>
<td>±.02</td>
<td>17</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Is easy to apply</td>
<td>.78</td>
<td>.90</td>
<td></td>
<td>±.12</td>
<td>18</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Is easy to understand</td>
<td>.78</td>
<td>.70</td>
<td></td>
<td>±.08</td>
<td>19</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Is independently metaevaluated</td>
<td>.78</td>
<td>.77</td>
<td></td>
<td>±.01</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Does not confuse grading, ranking, scoring, and apportioning</td>
<td>.78</td>
<td>.77</td>
<td></td>
<td>±.01</td>
<td>21</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Payoff is substantial</td>
<td>.77</td>
<td>.75</td>
<td></td>
<td>±.02</td>
<td>22</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Costs are reasonable for those submitting information</td>
<td>.66</td>
<td>.77</td>
<td></td>
<td>±.11</td>
<td>23</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Deals fairly with new as well as experienced researchers</td>
<td>.59</td>
<td>.68</td>
<td></td>
<td>±.09</td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Allows an appeal process</td>
<td>.49</td>
<td>.44</td>
<td></td>
<td>±.05</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
As shown in Table 30, the magnitude of the correlation coefficients between the two panels' consensus ratings on the five metadimensions on the same national models were moderate to large (i.e., $r = .78$ to $r = .98$), indicating highly consistent agreement between the two panels with respect to their ratings on the same pairs of countries.

Table 30

Pearson's Product-Moment Correlation Coefficients for Panels' Metadimensional Raw Scores

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity Raw Score</td>
<td>.90</td>
</tr>
<tr>
<td>Credibility Raw Score</td>
<td>.92</td>
</tr>
<tr>
<td>Utility Raw Score</td>
<td>.93</td>
</tr>
<tr>
<td>Cost-Effectiveness Raw Score</td>
<td>.84</td>
</tr>
<tr>
<td>Ethicality Raw Score</td>
<td>.78</td>
</tr>
<tr>
<td>Total Raw Score</td>
<td>.98</td>
</tr>
</tbody>
</table>

Note. * Each panel $N = 16$.

The correlation coefficient between the two panels' Stage II total weighted scores for the same countries was $r = .98$ ($df = 14, p < .01$) and the correlation coefficient between the Stage II rank ordering of total weighted scores was $r = .93$ ($df = 14, p < .01$) as shown in Table 31; the Spearman's rank correlation coefficient was $\rho = .93$.

Combined, these indices (i.e., reliability coefficients) reflect a relatively high degree of consistency (i.e., reliability) between the scores of the two panels. One conclusion that might follow, therefore, is that there is a high degree of agreement (i.e., interrater reliability) between the ratings of judge- and panel-pairs on the same countries.
### Table 31
**Total Weighted Scores and Rank Ordering of National Models**

<table>
<thead>
<tr>
<th>National Model</th>
<th>Total Weighted Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NZ</td>
<td>US</td>
</tr>
<tr>
<td>NZ</td>
<td>85.20</td>
<td>90.40</td>
</tr>
<tr>
<td>UK</td>
<td>73.20</td>
<td>70.00</td>
</tr>
<tr>
<td>NL</td>
<td>71.60</td>
<td>66.40</td>
</tr>
<tr>
<td>US</td>
<td>68.00</td>
<td>68.80</td>
</tr>
<tr>
<td>HK</td>
<td>57.60</td>
<td>51.20</td>
</tr>
<tr>
<td>AU</td>
<td>48.00</td>
<td>51.60</td>
</tr>
<tr>
<td>SE</td>
<td>46.20</td>
<td>42.00</td>
</tr>
<tr>
<td>PL</td>
<td>46.00</td>
<td>44.40</td>
</tr>
<tr>
<td>DE</td>
<td>45.80</td>
<td>46.80</td>
</tr>
<tr>
<td>JP</td>
<td>43.20</td>
<td>41.60</td>
</tr>
<tr>
<td>IE</td>
<td>41.60</td>
<td>38.40</td>
</tr>
<tr>
<td>BE</td>
<td>40.00</td>
<td>40.40</td>
</tr>
<tr>
<td>CZ</td>
<td>38.40</td>
<td>42.80</td>
</tr>
<tr>
<td>FI</td>
<td>36.40</td>
<td>30.80</td>
</tr>
<tr>
<td>HU</td>
<td>35.60</td>
<td>40.40</td>
</tr>
<tr>
<td>FR</td>
<td>10.80</td>
<td>11.20</td>
</tr>
</tbody>
</table>

*Note.* Each panel $N = 16$.

Cronbach’s coefficient $\alpha$ (reflecting the lower-bound of the estimate of reliability for the scores) for independent judges’ ratings are presented in Table 32 and for the two panels’ consensus ratings in Table 33.

As shown in Tables 32 and 33, coefficient $\alpha$ for independent NZ judges’ vis-à-vis independent US judges’ and the two panels’ consensus ratings in terms of the validity, credibility, utility, cost-effectiveness, and ethicality indicators were all within acceptable limits (i.e., $> .70$). Moreover, these estimates increased slightly from Stage I to Stage II.
Table 32
Cronbach's Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Independent Judges' Ratings

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>NZ $\alpha$</th>
<th>US $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity (5 indicators)</td>
<td>.95</td>
<td>.93</td>
</tr>
<tr>
<td>Credibility (5 indicators)</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>Utility (5 indicators)</td>
<td>.93</td>
<td>.93</td>
</tr>
<tr>
<td>Cost-Effectiveness (5 indicators)</td>
<td>.93</td>
<td>.94</td>
</tr>
<tr>
<td>Ethicality (5 indicators)</td>
<td>.88</td>
<td>.83</td>
</tr>
<tr>
<td>All Indices (25 indicators)</td>
<td>.98</td>
<td>.98</td>
</tr>
</tbody>
</table>

*Note.* *Each panel $N = 16.$

Table 33
Cronbach's Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Panels' Consensus Ratings

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>NZ $\alpha$</th>
<th>US $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity (5 indicators)</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>Credibility (5 indicators)</td>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>Utility (5 indicators)</td>
<td>.96</td>
<td>.94</td>
</tr>
<tr>
<td>Cost-Effectiveness (5 indicators)</td>
<td>.95</td>
<td>.94</td>
</tr>
<tr>
<td>Ethicality (5 indicators)</td>
<td>.97</td>
<td>.92</td>
</tr>
<tr>
<td>All Indices (25 indicators)</td>
<td>.99</td>
<td>.99</td>
</tr>
</tbody>
</table>

*Note.* *Each panel $N = 16.$
Reliability Estimates for Dyad-Pairs

The prior section presented various estimates of reliability using conventional methods. Given that the design of the study was dyadic (i.e., judges and panels were nonindependent; that is, they made judgments on the same objects and therefore cannot be treated as independent) it is necessary to estimate the correspondence between judges' and panels' ratings using alternative methods. Thus, ordinary correlation coefficients should not be used to measure nonindependence; rather, the correct estimate of correspondence is the intraclass correlation coefficient (Kenny, Kashy, & Cook, 2006). There are numerous methods for estimating intraclass correlations ($r_1$), but given the nature and type of data, the most appropriate for this study is the double-entry method. This method refers to the correlation coefficient between dyad-pairs as the pairwise correlation coefficient, symbolized as $r_p$. In this method, one judge or panel is designated as $X$, and the other as $Y$. Then the data are doubled, making each $X$ a $Y$, and each $Y$ an $X$. Thus, the sample size becomes $2n$, rather than $n$. The pairwise coefficient for the 'double-entered' data is then estimated as shown in Equation 10

$$r_p = \frac{SS_B - SS_W}{SS_B + SS_W}$$

(10)

Where $SS_B = df_B MS_B$ and $SS_W = df_w MS_w$. A pairwise correlation cannot be tested in the usual way and Griffin and Gonzalez (1995) recommend using $1/\sqrt{n}$ as the standard error for the test of $r_p$, resulting in a test statistic that is treated as a $Z$ statistic. The pairwise correlation coefficient is a maximum likelihood estimate of the correlation between the two panels' ratings, and these estimates are shown in Table 34. As the table shows, the pairwise intraclass correlation coefficient for panel dyad-pairs' validity raw scores was $r_p = .90 \ (p < .01)$, for credibility raw scores was $r_p = .91 \ (p < .01)$, for utility raw scores was $r_p = .93 \ (p < .01)$, for cost-effectiveness raw scores was $r_p = .83 \ (p <$
.01), for ethicality raw scores was $r_p = .76 (p < .01)$, and for total raw scores was $r_p = .98 (p < .01)$. These intraclass correlation coefficients provide a unique estimate of the relationship between dyad-pairs’ (i.e., judges of the same object) ratings on the same national model, and support the conclusion that the panels are nondistinguishable (i.e., nonindependent).

Table 34

Pairwise Intraclass Correlation Coefficients for Panels Dyad-Pairs’ Raw Scores

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>$r_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity Raw Score</td>
<td>.90</td>
</tr>
<tr>
<td>Credibility Raw Score</td>
<td>.91</td>
</tr>
<tr>
<td>Utility Raw Score</td>
<td>.93</td>
</tr>
<tr>
<td>Cost-Effectiveness Raw Score</td>
<td>.83</td>
</tr>
<tr>
<td>Ethicality Raw Score</td>
<td>.76</td>
</tr>
<tr>
<td>Total Raw Score</td>
<td>.98</td>
</tr>
</tbody>
</table>

* Double-entry listwise $N = 32$.

While these estimates do not differ markedly from the Pearson’s product-moment correlation coefficients between the two panels’ ratings on each of the metadimensions presented in the previous section (see Table 30), they do, however, indicate that the panels’ ratings are nonindependent and therefore should not be treated as independent samples or nested sets (e.g., panels within countries, or judges within panels within countries), for example. To do so would result in biased estimates (e.g., the pairwise intraclass correlation coefficients are generally smaller in magnitude than the product-moment correlation coefficients presented in the previous section), among others, and to treat dyadic data as if it were individual data would fail to optimize the conceptual design of the study (see Chapter III).
Likewise, Cronbach’s coefficient $\alpha$ can be estimated for the combined-panels’ ratings (i.e., listwise). As shown in Table 35, this procedure slightly improved the lower-bound estimates of reliability for independent panels’ scores as presented previously in Table 33.

Table 35

Cronbach’s Coefficient Alpha Estimates of the Lower-Bound Reliability of Scores for Combined-Panels’ Consensus Ratings

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity (5 indicators)</td>
<td>.95</td>
</tr>
<tr>
<td>Credibility (5 indicators)</td>
<td>.96</td>
</tr>
<tr>
<td>Utility (5 indicators)</td>
<td>.94</td>
</tr>
<tr>
<td>Cost-Effectiveness (5 indicators)</td>
<td>.94</td>
</tr>
<tr>
<td>Ethicality (5 indicators)</td>
<td>.95</td>
</tr>
<tr>
<td>All Indices (25 indicators)</td>
<td>.99</td>
</tr>
</tbody>
</table>

*Note.* Listwise $N = 32$.

Attributes Predicting Independent Judges’ “Overall Best Judgment”

Analysis of the attributes (i.e., metadimensional raw scores) predicting judge’s “overall best judgment” were also examined using elements of human judgment analysis, known as the Lens Model (see Chapter III), by regressing judge’s metadimensional raw scores on their “overall best judgment” (see Equation 9, p. 306) The intent of these analyses was to investigate which metadimensions were given the greatest weight in judge’s holistic appraisals of the national models.

For the NZ judges, the five attributes accounted for 98% (adjusted $R^2 = .98$) of the variability in their “overall best judgment” of the national models; $F(5, 10) = 126.90$,.
$p < .01$. As shown in Table 36, three of the five attributes were statistically significant predictors of NZ judges' "overall best judgment," with the greatest weight given to validity ($\beta = .53$), utility ($\beta = .32$), and ethicality ($\beta = .30$). Cost-effectiveness, however, was inversely associated with their overall holistic assessment of the national models ($\beta = -.09$); although, it is not a statistically significant ($p = .65$) predictor of their "overall best judgment."

Table 36

Summary of Regression Analysis for Attributes Predicting Independent NZ Judges' "Overall Best Judgment"

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$b$</th>
<th>$SEb$</th>
<th>$\beta$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.09</td>
<td>.03</td>
<td>.53</td>
<td>.01</td>
</tr>
<tr>
<td>Credibility</td>
<td>.00</td>
<td>.03</td>
<td>.02</td>
<td>.91</td>
</tr>
<tr>
<td>Utility</td>
<td>.06</td>
<td>.02</td>
<td>.32</td>
<td>.03</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>-.02</td>
<td>.04</td>
<td>-.09</td>
<td>.65</td>
</tr>
<tr>
<td>Ethicality</td>
<td>.06</td>
<td>.03</td>
<td>.30</td>
<td>.03</td>
</tr>
</tbody>
</table>

By contrast, the five attributes accounted for 98% (adjusted $R^2 = .98$) of the variability in US judges' "overall best judgment" of the national models; $F(5, 10) = 186.63, p < .01$. As shown in Table 37, four of the five attributes were statistically significant predictors of US judges' "overall best judgment," with the greatest weight given to validity ($\beta = .62$), followed by ethicality ($\beta = .27$), credibility ($\beta = .23$), and cost-effectiveness ($\beta = .21$). Utility, however, was inversely associated with their overall holistic assessment of the national models ($\beta = -.30$). Credibility, while not statistically significant ($p = .06$), did carry some weight ($\beta = .23$).
Table 37
Summary of Regression Analysis for Attributes Predicting Independent US Judges’ “Overall Best Judgment”

<table>
<thead>
<tr>
<th>Attribute</th>
<th>b</th>
<th>SE b</th>
<th>β</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.06</td>
<td>.01</td>
<td>.62</td>
<td>.00</td>
</tr>
<tr>
<td>Credibility</td>
<td>.02</td>
<td>.01</td>
<td>.23</td>
<td>.06</td>
</tr>
<tr>
<td>Utility</td>
<td>-.03</td>
<td>.01</td>
<td>-.30</td>
<td>.04</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>.02</td>
<td>.01</td>
<td>.21</td>
<td>.02</td>
</tr>
<tr>
<td>Ethicality</td>
<td>.03</td>
<td>.01</td>
<td>.27</td>
<td>.02</td>
</tr>
</tbody>
</table>

Combined, the five attributes accounted for 97% (adjusted $R^2 = .97$) of the variability in NZ and US judges’ collective “overall best judgment” of the national models; $F(5, 26) = 175.18, p < .01$. As shown in Table 38, two of the five attributes were statistically significant predictors of the two panels’ judges “overall best judgment,” with the greatest weight given to validity ($\beta = .44$) and ethicality ($\beta = .23$).

Table 38
Summary of Regression Analysis for Attributes Predicting Combined Judges’ “Overall Best Judgment”

<table>
<thead>
<tr>
<th>Attribute</th>
<th>b</th>
<th>SE b</th>
<th>β</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.08</td>
<td>.02</td>
<td>.44</td>
<td>.00</td>
</tr>
<tr>
<td>Credibility</td>
<td>.03</td>
<td>.02</td>
<td>.14</td>
<td>.22</td>
</tr>
<tr>
<td>Utility</td>
<td>.01</td>
<td>.02</td>
<td>.07</td>
<td>.52</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>.03</td>
<td>.02</td>
<td>.16</td>
<td>.11</td>
</tr>
<tr>
<td>Ethicality</td>
<td>.05</td>
<td>.02</td>
<td>.23</td>
<td>.01</td>
</tr>
</tbody>
</table>

Table 39 presents a comparison of the standardized beta coefficients (i.e., $\beta$s) for the attributes predicting independent NZ and US judges’ and combined-panels’ judges “overall best judgments” of the national models.
Table 39

Comparison of Standardized Beta Weights for Attributes Predicting NZ, US, and Combined Judges’ “Overall Best Judgment”

<table>
<thead>
<tr>
<th></th>
<th>NZ β</th>
<th>US β</th>
<th>Combined β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>.53</td>
<td>.62</td>
<td>.44</td>
</tr>
<tr>
<td>Credibility</td>
<td>.02</td>
<td>.23</td>
<td>.14</td>
</tr>
<tr>
<td>Utility</td>
<td>.32</td>
<td>-.30</td>
<td>.07</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>-.09</td>
<td>.21</td>
<td>.16</td>
</tr>
<tr>
<td>Ethicality</td>
<td>.30</td>
<td>.27</td>
<td>.23</td>
</tr>
</tbody>
</table>

Feedback from both panels’ judges following the two-stage rating procedure support these conclusions in that judges indicated nearly unanimously that validity is by far the most important attribute of a good national research evaluation model. For example, as one US judge stated:

If it is not valid then the rest do not really matter...it can have credibility, have utility, be cost-effective, and be ethical, and still not be valid...and if it is not valid, then how can any decisions based on it be rational?

Generalizability of Judges’ Ratings

A single facet, design 4 (i.e., each national model was rated by more than one judge; there are different judges for each national model) generalizability (G) study was conducted to examine the reliability of the results of the study if conducted under the same or similar conditions (namely, generalizability). Despite the training and calibration procedures, judges were specified as a facet (i.e., source of variation) in the G study because variability in the obtained scores due to judge differences would be undesirable and contribute to unreliability of the results. The formula used to estimate the generalizability coefficient is given in Equation 11
\[ \hat{\rho}_i^2 = \frac{\sigma^2_p}{\sigma^2_p + (\sigma^2_i + \sigma^2_e)/n_i} \]  

(11)

where the \( I \) subscript on \( \hat{\rho}_i^2 \) is the average rating (Crocker & Algina, 1986). Or, directly from the mean squares as given in Equation 12

\[ \hat{\rho}_i^2 = \frac{MS_p - MS_r}{MS_p + n_i MS_i / n_p n_i' + (n_p n_i' - n_p n_i - n_i) MS_e / n_p n_i'} \]  

(12)

Thus, the use of multiple judges should reduce the effects of judge variance and residual variance on the observed score variance and increase the generalizability coefficient.

The variance estimates for the G study were derived from the results of a two-factor analysis of variance (ANOVA), using the data presented in Table 40, classified in terms of two dimensions (i.e., factors): judges and national models. The ANOVA yielded three distinct sources of variance necessary for estimating the generalizability coefficient: judges, national models, and judges \( \times \) national models. In addition, the interaction term, judges \( \times \) national models, contains several sources of variance that cannot be separated, including information about random, unaccounted for variance (sometimes referred to as the residual error term).

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judges</td>
<td>1643.00</td>
<td>6</td>
<td>273.83</td>
</tr>
<tr>
<td>National Models</td>
<td>225842.00</td>
<td>10</td>
<td>22584.20</td>
</tr>
<tr>
<td>Judges ( \times ) National Models</td>
<td>1194.00</td>
<td>10</td>
<td>119.40</td>
</tr>
</tbody>
</table>
The sources of variance refer to different, systematic ways that the scores can vary (Brennen, 1983). The judge component indicates systematic and overall differences in the way that judges rate the models. If one judge simply rated consistently higher than another judge then that would show in the judge component of variance. The national models component reflects real differences in the merits of the models. The judges \times national models component indicates that the relative merits of the models is different for the judges.

The sources of variance presented in Table 40 yielded a generalizability coefficient of $\phi^2 = .99$. The percentages of variance in scores that are attributable to the different sources of variance are presented in Table 41. As shown in the table, 94% of the variability in scores can be attributed to the national models (this variability is desirable), whereas less than 1% can be attributed to judges (e.g., systematic differences in ratings, which is undesirable) and 6% to error, or the interaction between judges and objects of judgment (i.e., national models).

Table 41

Variance Accounted for by Components

<table>
<thead>
<tr>
<th>Component</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judges</td>
<td>0.48%</td>
</tr>
<tr>
<td>National Models</td>
<td>93.55%</td>
</tr>
<tr>
<td>Judges × National Models</td>
<td>5.97%</td>
</tr>
</tbody>
</table>
Classification of Funding System Archetypes

A descriptive discriminate function analysis (DFA) was used to assess the degree to which the five metadimensions discriminated between Type I, Type II, and Type III funding system archetypes (see Chapter II). It was determined that the homogeneity of variance assumption was met (Box’s $M$ \( F(30, 1773.92) = 1.73, p = .01 \) = 71.94), indicating that covariance matrices could be pooled for the analysis. There was a large canonical correlation ($R_e = .84$) on Function 1 with an effect size of $R^2_e = 70.39\%$ between the grouping variable (Type I, Type II, and Type III) and the composite predictor variables (validity, credibility, utility, cost-effectiveness, and ethicality). The full model test for the function was significant; $\Lambda = .28$, $X^2(10, 32) = 34.49$, $p < .01$. However, as shown in Table 42, the test of Function 2 (i.e., discrimination between Type II and Type III models) was not significant and therefore excluded from subsequent analyses. The means and standard deviations for each of the three types of models on the five metadimensions are presented in Table 43.

Table 42

<table>
<thead>
<tr>
<th>Function</th>
<th>$\Lambda$</th>
<th>$X^2$</th>
<th>$df$</th>
<th>$p$</th>
<th>$R_e$</th>
<th>$R^2_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>.279</td>
<td>34.49</td>
<td>10</td>
<td>.00</td>
<td>.84</td>
<td>70.39%</td>
</tr>
<tr>
<td>2</td>
<td>.940</td>
<td>1.68</td>
<td>4</td>
<td>.79</td>
<td>.25</td>
<td>6.05%</td>
</tr>
</tbody>
</table>
Table 43
Means and Standard Deviations on the Five Metadimensions for the Three Funding System Archetypes

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>Type I</th>
<th></th>
<th>Type II</th>
<th></th>
<th>Type III</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>73.00</td>
<td>11.11</td>
<td>41.35</td>
<td>8.12</td>
<td>36.50</td>
<td>21.21</td>
</tr>
<tr>
<td>Credibility</td>
<td>74.00</td>
<td>12.40</td>
<td>44.07</td>
<td>5.69</td>
<td>37.00</td>
<td>17.59</td>
</tr>
<tr>
<td>Utility</td>
<td>71.40</td>
<td>10.45</td>
<td>42.57</td>
<td>4.92</td>
<td>33.50</td>
<td>20.77</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>68.60</td>
<td>13.92</td>
<td>41.85</td>
<td>7.37</td>
<td>34.25</td>
<td>15.02</td>
</tr>
<tr>
<td>Ethicality</td>
<td>63.80</td>
<td>14.97</td>
<td>38.42</td>
<td>8.98</td>
<td>33.25</td>
<td>11.94</td>
</tr>
</tbody>
</table>

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

Standardized discriminant function coefficients and structure coefficients were examined to determine which of the metadimensions contributed to the differences in the three types of models. As shown in Table 44, validity emerges as the metadimension most correlated with the grouping variable (i.e., type of model) on Function 1, meaning that it contributes the most to separation of the models.

Table 44
Standardized Discriminant Function and Structure Coefficients

<table>
<thead>
<tr>
<th>Metadimension</th>
<th>Coefficient</th>
<th>$r_s$</th>
<th>$r^2_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>-.31</td>
<td>.89</td>
<td>79.74%</td>
</tr>
<tr>
<td>Credibility</td>
<td>.90</td>
<td>.86</td>
<td>73.27%</td>
</tr>
<tr>
<td>Utility</td>
<td>-.07</td>
<td>.81</td>
<td>65.77%</td>
</tr>
<tr>
<td>Cost-Effectiveness</td>
<td>.55</td>
<td>.81</td>
<td>65.12%</td>
</tr>
<tr>
<td>Ethicality</td>
<td>.09</td>
<td>.74</td>
<td>54.16%</td>
</tr>
</tbody>
</table>

Note. *Structure coefficients ($r_s$) may range from -1 to 1, and there is no shared variance between the predictor variables (i.e., metadimensions).
As the group centroids show (see Table 45), Type I models were substantially higher on the composite metadimensions than Type II and Type III models. This and the structure coefficients indicate that the differences (i.e., separation) observed on Function 1 can be attributed mostly to validity, and to some extent credibility, utility, cost-effectiveness, and ethicality given that these were all positively correlated in the function (see Table 44). Therefore, Type I models have more of these traits (validity, credibility, utility, cost-effectiveness, and ethicality) than either Type II or Type III models in the linear equation.

Table 45
Group Centroids

<table>
<thead>
<tr>
<th>Model</th>
<th>Function 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>2.13</td>
</tr>
<tr>
<td>Type II</td>
<td>-0.69</td>
</tr>
<tr>
<td>Type III</td>
<td>-1.47</td>
</tr>
</tbody>
</table>

Summary of Key Findings

The results of this study clearly demonstrate that New Zealand's PBRF is considered the highest quality model for evaluating research and allocating research funding in comparison to the other 15 national models included in the sample (for the NZ panel +12 over the nearest competitor and for the US panel +20 over the nearest competitor; the United Kingdom's RAE in both cases). The strength of the PBRF, according to the panel judges, lies in its unit of analysis (individuals versus institutions, programs, or disciplines, for example), its comprehensiveness (e.g., of criteria), its transparency (e.g., of procedures and guiding principles), its overall approach (i.e., mixed
in terms of expert panel review and quantitative indicators of research quality), that differential funding is allocated according to the system's research quality categories, the consideration given to new and emerging researchers, and that the exercise is independently metaevaluated. Despite its relatively high rating, the model also has several weaknesses. For example, judges indicated that the system is overly complex, expensive in terms of monetary and non-monetary costs (e.g., time, expertise required), and that although individuals may be the unit of analysis, funding is distributed at an institutional level.

By contrast, the French model (despite the establishment of the National Committee for Evaluation in 1984 and the National Committee for Research Evaluation in 1989) was rated very low on all five dimensions by both panels' judges. These negative ratings were in part due to the separation of assessment and decision structures, lack of transparency, clashes of interest in the French peer review system, resistance to evaluation in the research community (e.g., in 2004, French researchers began a protest movement which sought to inform the government and their fellow citizens on the negative consequences of dysfunctions in the French research system), the promotion of researchers by seniority rather than quality, and the overemphasis on self-evaluation in the French national model, according to the panels' judges.
CHAPTER VII

CONCLUSIONS

This final chapter consists of four sections. First, the central findings from the comparative evaluation of the 16 national research evaluation models are discussed. Second, the limitations of the study of the national models are given. Third, a short discussion of potential implications for research evaluation theory, practice, and policy is given. Finally, prospective areas of additional research or future work, simply titled 'visions for the future,' are enumerated.

Central Findings

More than two-thirds of the models were assigned to a quality category of F (i.e., absence of merit; clearly inadequate; fatal deficiencies). The various systems in place are unlikely to be dismantled anytime soon and demands for accountability will not go away. It may be that research evaluation that does not work is worse than none at all, and it costs a substantial amount of time, money, and expertise, as well as exhausting and sometimes demoralizing the researchers obliged to participate (Nature, 2006).

The United Kingdom led the field of large-scale evaluations of research with the introduction of the RAE in 1986, which eventually lead to the introduction of similar systems throughout the world. But after the 2008 RAE, the United Kingdom government will move to a simpler, more cost-effective metrics-based system for assessing research quality and allocating funding. The most favored model, New
Zealand's PBRF, also has had several concerns raised (Curtis & Matthewman, 2005), such as the real cost-benefit ratio of participation, with reports that many universities have spent more on the exercise than they will gain in funding increases (Nature, 2006). Questions have also arisen as to whether the quality of research has improved as a direct result of the assessment. The PBRF unit of analysis (the individual) has received the most criticism and, after this year's assessment, this matter will be reviewed. In Australia, critics of the RQF suggest that it might be more cost-effective to modify existing research assessment processes than to undertake a new and potentially costly and arduous exercise (Donovan, 2006, November 1; Shewan & Coats, 2006). However, proponents of the RQF have concluded that it "will provide the first national system-wide evaluation of the public value of research (or research 'impact'), which, along with the measurement of research 'quality', will inform the distribution of national research funding" (Donovan, 2008). While the United States is generally regarded as having some of the strongest research performers and performances in the world, it has not been the most innovative with regard to its research evaluation systems (Dill, 2003, March). Furthermore, the process (e.g., GPRA, OMB's PART) contains little of the transparency Americans expect of their government.

Notably absent in most national systems (the Netherlands, New Zealand, and to a lesser extent the United Kingdom and United States, being exceptions) are serious independent metaevaluations of government-performed evaluations of research (whether these evaluation are conducted by internal, external, or via mixed expert panels or using other methods; see Chapters I and II). Since its introduction in the 1960s (Scriven, 1969), metaevaluation (the evaluation of evaluations) has become the standard for assessing the quality of evaluations (Cooksy & Caracelli, 2005), and has even been referred to as an 'imperative' or professional obligation (Stufflebeam, 2001). Originally envisioned as a
procedure for assuring the quality of evaluations (some have described metaevaluation as an auditing process), metaevaluation has evolved to include other potential applications, including for example, giving credibility to the public (i.e., taxpayers) for the manner in which research funds are distributed, improving future assessments, and determining the degree to which an evaluation's procedures, conclusions, and recommendations (if any) are reasonable, credible, and justifiable. In addition, metaevaluation can be done both formatively, to improve the design or process of an evaluation in-progress, and summatively, in order to meet demands for accountability (Davidson, 2005; Greene, 1992; Scriven, 1991). The quality criteria applied to evaluations of evaluations, however, are highly contestable and lack any consensus. For example, the Joint Committee on Standards for Educational Evaluation (1994) suggests standards of utility, feasibility, propriety, and accuracy. Other potential criteria include for example (Cooksy & Caracelli, 2005), transparency (Henry, 2001), balance (Patton, 1997), relevance (Patton, 1997), credibility (Guba & Lincoln, 1989), validity (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), legitimacy (Schwartz & Mayne, 2004), cultural competence (Kirkhart, 2004), and systematic (Scriven, 1991; Weiss, 1998).

Another major problem that plagues a large majority of national research evaluation systems is the tendency to treat past performance as the sole indicator of future performance. That is, to treat a problem in predictive evaluation (i.e., ex ante) as essentially solved by doing retrospective evaluation (i.e., ex post). This is hazardous, and calls attention to the fact that most countries do not usually distinguish the two sufficiently well (Scriven, 2007, May 7). In part, these tendencies arise from the failure to adequately distinguish clearly between the evaluation of research and the funding of research and fundamentally these systems are designed to reward good research performance and to be punitive against poor research performance, whether the
reasoning and logic used to reach these conclusions is faulty or not (in part, high-quality independent metaevaluation could serve to verify the extent to which evaluative conclusions about where funding should be distributed were valid).

On the surface, the evaluation of government-financed research appears quite well structured with its extensive literature on its objects, motives, methods and procedures, criteria and standards, difficulties, and results (OECD, 1997; Giorgi & Tandon, 2000; Jeannin & Devillard, 2005). Simultaneously, most governments around the world recognize that current methods are not sufficient for current needs, and are now funding efforts to find new and improved methods for evaluating research. Nevertheless, evaluation is not yet a major factor in many national research policy systems and has had little, if any, influence on policy decisions; often due to the tenuous link between policy- and decision-makers and policy analysts or evaluators. Thus, insights, even when they exist, may not even make it into the decision-making process. More often than not, the allocation of research funding is a political game in which financial apportionments are determined by interest groups competing for government resources by demonstrating the importance of their research; or, funds are earmarked by decision-makers according to their own agendas or priority areas. Therefore, it seems unlikely that there is a 'one size fits all' model for evaluating research or distributing research funds at a national-level given the large number of contextually-dependent variables to be considered (e.g., national needs, value systems, resources, priority areas, agendas). Nonetheless, it is equally important to give due consideration to the notion that while large-scale assessment exercises such as those used by Hong Kong, the Netherlands, New Zealand, and the United Kingdom may not be a 'fits all' solution (e.g., some suggest that the sheer geographic size of the United States, for example, prohibits the application of such systems), that the bulk funding and indicator-driven types of
mechanisms failed to meet even the lowest quality standard thresholds.

In any case, each of the countries included in the study should be commended for at least having minimal structures in place for systematically and strategically evaluating government-funded research. That most national models did not meet the minimal criterion-referenced standard for being placed in a higher quality category should not be taken as a sign of failure, but rather as an impetus for making improvements or modifications to existing principles and procedures.

**Limitations**

A major, albeit anticipated, limitation to the comparative evaluation of the national models for evaluating research was how national policies were summarized and represented. That is, the narratives judges used in their appraisals of the national models presented national policies and principles and practices for evaluating government-financed research as they were intended or envisioned, not necessarily how they are put into practice (i.e., fidelity). However, two-thirds of the narratives were verified by policymakers and/or scholars of research evaluation within the respective countries, explicitly in reference to the five merit-defining metadimensions used to assess the quality of the national models. A second, equally relevant, limitation to the study of national models was the selection of sites and judges. On the one hand, selection was purposeful but also one of convenience, under consideration of other relevant and demonstrable constraints (e.g., resources, language). On the other, it is quite likely that if the panels were conducted in one or more of the European or Asian countries that a different set of personal and/or cultural values would have been in operation and that the final results may have differed to some extent. Finally, and not unrelated to the former, is the notion of whose values (i.e., merit-defining criteria) should be used to determine what
constitutes a good national model for evaluating research. But, the criteria used (i.e., validity, credibility, utility, cost-effectiveness, and ethicality) are aimed to convince most professional evaluators and researchers as highly plausible.

Implications

The implications emerging from this research are multifaceted, including several for research evaluation theory and practice as well as for research policy. In the following, a partial, but by no means complete, list of these implications is presented.

Theory and Practice

While the study of national models does not contribute greatly to the theory and practice of evaluating researchers and their research, the work presented in this dissertation does demonstrate that such evaluations are much more than merely 'research on research,' and that they can be conducted systematically and objectively. It, like all forms of evaluation, is governed by the same fundamental reasoning and logic which can be reduced to matters such as its definition (i.e., systematically and objectively determining the merit, worth, or significance of things) and the definition of its major concepts (e.g., grading, ranking, scoring, apportioning, synthesis, barring, steeping, scaling, profiling), the nature of its relations to other subjects and other disciplines (i.e., the transdisciplinary model), and the rules of inference that govern it (e.g., beyond reasonable doubt; the balance or weight of evidence). In a like manner, evaluations of researchers and their research, whether ex ante or ex post, require the same application of the same core dimensions of evaluation (i.e., process evaluation, outcomes evaluation, costs evaluation, comparative evaluation, and generalizability evaluation) that are necessary for most other evaluands, and which apply across the major fields of
evaluation (i.e., program evaluation, personnel evaluation, performance evaluation, policy evaluation, portfolio evaluation, and proposal evaluation).

The importance of these fundamental principles should not be underestimated—nor should the distinction between indicators and criteria, values and valuing (including the sources of relevant and demonstrable values), a complete understanding of the basic evaluative predicates (i.e., merit, worth, and significance), the tripartite taxonomy (i.e., formative, summative, and ascriptive), types of evaluative claims (i.e., personal preference claims, market value claims, contextual value claims, and essentially evaluative claims), and types of premises (i.e., factual and value)—in understanding how evaluation is different from other undertakings, and what very specific procedures are required to respond to truly evaluative questions. While this work draws extensively from the field’s pioneers (Scriven, in particular), it is hoped that those unfamiliar with the basic logic and methodology of evaluation can draw on that which is truly evaluation specific.

**Policy**

The rigorous evaluation of research projects and programmes is in increasingly common demand across the world. Attempts have been made to implement it in Europe, Japan and the United States—but until the calibre of these efforts improves, scientists will continue, justifiably, to view them with suspicion...[and]...policy-makers have talked for years about the need to rigorously evaluate research programmes that consume billions of dollars of taxpayers’ money. Researchers—especially those doing basic research that can’t be readily tied to concrete outcomes—have tended to be sceptical. Nonetheless, evaluation is now underway on a significant scale in every major economy (Nature, 2006, p. 1).

In the United States alone, as evidenced by the concerns of the American Council on Competitiveness, the initiative of the White House Office of Science and Technology Policy (OSTP), and now NSF’s funding of studies to build a science of science and innovation policy, and similar efforts throughout Europe, in Japan, and in
Korea, a primary interest is assessment to understand how to improve research so that it can effectively contribute to national goals, as well as prescribing evidence-based evaluation methodologies and deterring the controversy surrounding the issue of 'scientific' methodology in federally-funded research (Julnes & Rog, 2007). This research should provide an important baseline characterization of various national research evaluation systems and mechanisms for evaluators, research managers, and policy makers.

The disagreement surrounding the issue of scientific methodology in federally-funded research, however, remains on the immediate horizon. Furthermore, panels awarding up to 20 additional points for random assignment experimental designs (the so-called '20% solution') have also been criticized for tying the hands of reviewers who feel obligated to recommend funding mediocre experimental design projects over well-designed non-experimental projects. These controversies are but part of a larger debate about how research should be evaluated by governments and, as noted, there has been an evolution in the procedures used for evaluating government-funded research, but these changes have not been matched by a consensus on how research should be evaluated.

The focus of attention on evidence of one form or another, and ideological debates over research evaluation policies, should bring about a better balance of values (e.g., more valid, credible, and useful research is funded and done so on the basis of procedures that are ethical and less costly). Those involved in the design and execution of research funding allocation decisions should consider these notions carefully as it could provide the basis for improving evaluation methods and thereby making better use of resources devoted to research. While the major characteristics may be clear enough, the details get more complicated when looking either across or within government
agencies and a policy that yields quality research on human stem cells might be less effective in promoting quality research in education.

Finally, the influence of evaluation on policy is generally conceived of in three broad categories (Caracelli & Preskill, 2000; Cousins & Leithwood, 1986); although it can also be completely ignored. First, evaluation can be used instrumentally in policy; that is, directly for decision making to give direction to policy and practice. Second, evaluation can be used conceptually; that is, evaluation does not always directly influence decisions or actions, but rather evaluation sometimes acts in a more indirect, subtle way to provide generalizations, ideas, or concepts that influence policy decisions. Third, evaluation can be used politically or symbolically; that is, to provide legitimization or to justify preexisting preferences for a decision or action. Also, evaluation can have what has been called “imposed use” (Weiss, Murphy-Graham, & Birkeland, 2005, p. 25). Imposed use occurs when demands for specific action are forced from more powerful operating levels (e.g., federal government) on less powerful operating levels (e.g., state government) on the basis of evaluation results.

Reforming the Evaluation of Research: Visions for the Future

Improving the quality of research, and the evaluations of them, is more important than ever. In theory, good evaluations of research should both increase the quality of what is done and decrease the cost of doing it. In this final section, a brief list of ‘visions for the future’ for reforming the evaluation of research is set forth. These points are in no particular order of importance and are intended only as an appeal for future study and consideration.
That Which can be Counted...Should it?

Not everything that can be counted counts; not everything that counts can be counted (Einstein, 1879-1955).

The appeal of ‘countable’ indicators of research quality is undeniable. Such techniques are relatively inexpensive, quick, and easy, and many of the countries placing a greater emphasis on public accountability for government research funding are starting to use quantitative performance indicators for the distribution of government research funds. For example, in Australian universities the use of quantitative formulas to allocate the research component of university block grants to institutions has been in place for a decade, and thus the system provides fertile ground for using bibliometrics to examine the effects of such policies on academic output. An analysis of Australian data from ISI’s major citation indexes demonstrates the academic response to the linking of funds, at least in part, to productivity measures undifferentiated by any measure of quality—publication numbers jumped dramatically, with the highest percentage increase in lower impact journals (Butler, 2005). It can be reasonably inferred, then, that while the notion of quick and easy methods for assessing quality (or quantity) is an attractive one, that such systems are not able to adequately deal with or account for ‘game playing’ (i.e., researchers manipulating the system for their own benefit).

The Research-Teaching Relationship Reconsidered

Teaching and research have traditionally been regarded as two independent tasks or duties; that is, “different enterprises” (Hattie & Marsh, 1996, p. 513). But, in evaluating a nation’s research, the teaching component is not entirely irrelevant for one simple reason: Without it, the supply of future researchers is jeopardized. In fact, a brain drain or human capital flight (i.e., an emigration of trained and talented individuals to
other nations or jurisdictions, due to conflicts, or lack of opportunity, where they are living) is quite likely when research evaluation systems do not give careful consideration or attention to new and emerging researchers. Its counterpart is brain gain in the areas to which talent migrates. Brain drain can occur either when individuals who study abroad and complete their education do not return to their home country, or when individuals educated in their home country emigrate for higher wages or better opportunities (Beine, Docquier, & Rapoport, 2001). The second form is arguably worse, because it drains more resources from the home country. This phenomenon is perhaps most problematic for developing nations, where it is widespread (OECD, 2003b). In these countries, higher education and professional certification are often viewed as the surest path to escape from a troubled economy or difficult political situation. Thus, investments in new and emerging researchers, or students in research programs, are an investment in a nation's economic and intellectual security.

These concerns are especially profound for Poland (since its introduction into the EU at least one-million Polish people, usually young and educated, have emigrated to Western European; mostly to the United Kingdom and Ireland), for France (with young graduates moving to Britain, the United States, and Canada because of economic and labor regulations making it increasingly difficult for new graduate to find employment), and New Zealand (which is experiencing an economic brain drain for a variety of reasons, with Australia being the main beneficiary; in 2005, nearly 25% of all New Zealand-born people with tertiary educations lived overseas).319

---

319 Historically a large proportion of New Zealand youth have always traveled overseas on Overseas Experiences (OEs). However, the vast majority would return home to start careers and families in New Zealand. In recent times however, the number of émigrés choosing to remain expatriate (mostly graduates of higher education) has steadily increased.
The relative success of fostering return migration in Chinese Taipei, Korea, and Ireland has been attributed to the opening of their economies and policies to foster domestic investments in innovation and R&D. Developing countries with an R&D infrastructure, like India, are more likely to attract the return of migrants. ‘Scientific diaspora’ and ‘immigrant entrepreneur networks’ can also help sending countries capture benefits and know-how from emigrants overseas (Cervantes & Guellec, 2002, May). Grass roots initiatives in South Africa and Latin America have been developed to link researchers abroad to networks in their home countries and Indian professionals in the United States have been the primary drivers of knowledge and capital flows to India. The Indian government has contributed to the emergence of these private networks through legislative and tax rules that encourage remittances and investment from Indians abroad. The diaspora idea has been put to work by advanced countries too, like Switzerland, whose online network, Swiss-List.com was established to encourage networking among Swiss scientists in the United States and to foster contacts with peers in Switzerland (Cervantes & Guellec, 2002, May).

The Role of Theory: A Cynical View

I thought I’d get your theories, mock them, and then embrace my own (Friedman & Lo, 2006).

Many, many evaluators of research, R&D, S&T, and RTDI argue for a greater role of theory in evaluating research (e.g., Arnold, 2004; Molas-Gallart & Davies, 2006), particularly in reference to research portfolios (Jordan, Hage, & Mote, 2006, November) and complex systems approaches (Byrne, 1997; Gleick, 1988; Kahn & Mann, 1957; von Bertalanffy, 1968; Williams & Imam, 2006). However, hinging the entire endeavor on

---

320 Theory as applied to evaluations of research are usually of the type such as identifying the causal chain connecting policy interventions and outcomes including eventual societal impact.
a theory, even if the merits of the theory can be demonstrated, is sometimes suicidal (Scriven, 1997). This is true for two reasons. First, theory does not resolve the fundamental questions. That is, is the research meritorious or valuable? Second, theory is useful to scientists and researchers but not research managers, policy-makers, or other decision-makers. In rare instances, however, a theory can be useful for evaluating research. For example, theory can play a useful role for evaluating a research plan or proposed research (e.g., propositions such as, if \( x \) then \( y \) or if more \( x \) then more \( y \)). Ultimately, the role of theory in the evaluation of research is the difference between that of an academic endeavor and that of a pragmatic one.

**Values Duality: The Tension between Competing Demands**

Scientific research and evaluation share a number of common characteristics, and it stands to reason that they have more commonalities than differences. Nevertheless, one thing is clear, while (most) evaluators actively attend to value and values (i.e., 'values-engaged;' Greene, DeStefano, Burgon, & Hall, 2006), (most) scientific researchers, by contrast, actively avoid them—as to contaminate oneself with values would be to thwart objectivity. In science, the ideal of objectivity is generally considered to come about as a result of strict observance of the scientific method and is intimately related to the aim of reproducibility (i.e., value-neutral).

Matters get increasingly complicated when one considers that evaluation also has its own set of values to contend with; many of which are at odds with one another. There are five (probably many more) pairs of values that frequently compete with one another in most types of research evaluation. Hackett (1997) outlined several of these nearly a decade ago, and the principal value-pairs to be considered (and dealt with) are:
Effectiveness versus efficiency. On the one hand, evaluation should be capable of discriminating the meritorious from the meritless, the worthwhile from the worthless, and the important from the trivial. However, in terms of effectiveness, there is also a tension between sensitivity and selectiveness (i.e., in statistical terminology Type I and Type II errors). Occasionally an evaluation will make a false-positive or false-negative (e.g., a poor piece of research is rated highly and excellent research goes unnoticed). On the other hand, evaluation should be low-cost (i.e., efficient). That is, good evaluation is economical in terms of the benefits produced by it. However, increases in one often result in sacrifices to the other.

Meritocratic versus fair. Evaluation should confer judgments according to strict criteria or dimensions of merit. However, evaluation is also expected to give due attention to fairness, which may underemphasize merit. Merit may conflict with fairness when sometimes irrelevant or idiosyncratic factors (e.g., age, race/ethnicity, gender) or “...political considerations...[or]...priorities enter into a decision” (Hackett, 1997, p. 58), for instance.

Reliability versus validity. Ideally, evaluation is to be simultaneously reliable and valid—measuring merit (or worth or significance) with little random or systematic error. In practice, however, one is often sacrificed at the expense of the other. That is, “narrow, rigid, quantifiable criteria may contribute to reliability (because they can be applied again and again...with quite consistent results)” (Hackett, 1997, p. 58), but these criteria may not accurately reflect the true merit or worth of research—that is, they might have low validity.

Accountability versus autonomy. In evaluating researchers, they should be held accountable to their peers, their funders, the public, and other relevant stakeholders. At the same time, they should be afforded some degree of autonomy from public scrutiny.

Responsive versus inertial. Evaluations of research should be responsive to new ideas and needs and translate those needs into scientific priorities, but at the same time impart stability and continuity to the scientific enterprise. This is the essential tension between tradition and originality.

Flexibility versus rigidity. On the one hand, evaluation should be flexible or responsive in terms of criteria, questions, design, methodology, and so forth (Patton, 1997; Stake, 2004). On the other hand, it is often claimed that evaluation should be conducted within (i.e., paradigm) a sometimes rigid theoretical or methodological framework (e.g., theory-driven or theory-based, randomized controlled trials [RCTs], goal-based or goal-achievement). Thus, rigidity may conflict with flexibility when the
evaluation pursues methodological rigor over needs (e.g., answering or responding to stakeholders' key questions, including all relevant and demonstrable stakeholder values), for example.

**Independence versus inclusiveness.** Independence, though only a relative notion, is widely perceived as the ideal for reducing some forms of bias and maintaining neutrality. By contrast, inclusiveness, inherent in participatory and collaborative evaluation methodologies (Cousins & Earl, 1992; Cousins & Whitmore, 1998; Lennie, 2006; Rodriguez-Campos, 2005), are intended to increase buy-in, usefulness, meaningfulness, and value (Coryn, 2006b), yet often at the cost of (external) credibility and impartiality. The value dilemma is one that essentially exist on a continuum from completely independent, external evaluation at one end to various dependent, internal evaluation methodologies (e.g., Fetterman's [1994, 2001] empowerment evaluation; appreciative inquiry [Preskill & Coghlan, 2004]) at the other.

**Risk versus uncertainty.** Ex ante evaluations of research can be characterized as relatively low in terms of uncertainty (assuming they are valid), but high in terms of risk. For instance, will the research produce reasonable social, economic, or other types of benefits? Is the researcher, research group, institution, or program capable of conducting the research? Is the necessary infrastructure in place for conducting the research? Do the potential benefits of the research outweigh the costs? What other research could be funded instead (i.e., opportunity costs)? By contrast, ex post evaluations of research are highly uncertain (i.e., actual social, economic, or other benefits), yet low-risk.

As should be evident from the above, many of the value-pairs are not discrete and in some cases they are orthogonal. Furthermore, the list is not intended to be exhaustive and does not include other opposing pairs such as those often encountered between stakeholder groups (e.g., divergent values between funders of research and researchers themselves or end-users), for example.

**The Trifecta: Type A, B, and C Evaluations of Research**

Scriven recently noted that evaluation of research can occur in three main practical contexts, each of which requires very different approaches (M. Scriven, personal communication, December 15, 2006). Therefore, it is important to distinguish
clearly between:

Type A  Retrospective evaluation of research (e.g., for making a research award)
Type B  Prospective evaluation of research (e.g., for hiring a researcher)
Type C  Evaluation of research investment (e.g., for funding a set of proposals).

According to Scriven, these distinctions are often glossed over because B involves A, and C involves B; but in each case, the later task in the sequence involves major further difficulties. Moreover, even A, the simplest task, which is correctly regarded as the main domain of peer review, involves many difficulties as commonly practiced that are often not faced as together constituting a severe limitation on validity. Also, B and C can not be ‘solved’ in the sense of dealt with via an algorithm, let alone by peer review as usually understood.

Funding Research versus Evaluating Research: Getting it Right

Again, Scriven (2006f) has reconciled one of the fundamental errors often made in evaluating research; that is, the difference between evaluating it and funding it:

The evaluation of research and researchers is an example of a fairly basic kind of evaluation...and these are tasks that we know a good deal about doing...But the evaluation of research funding is another kind of animal altogether. It aims for an apportionment or allocation decision, which is either something essentially different from evaluation or, with a stretch, a highly complex kind of evaluation decision. It is certainly a decision that depends on more than one kind of basic evaluation, but it depends on them in a way that has never been reduced to a formula or computer program. To a substantial degree, the major efforts that are being made by a number of countries to allocate governmental research funding to researchers and research projects in a rational way have been confounded by a failure to make this distinction. They tend to suggest, and be attacked as if this suggestion is correct, that they are engaged in the evaluation of research, when in fact they are, quite rightly, engaged in
working out the best way to allocate research funds (pp. 120-121).

These errors are particularly evident in countries using block funding or indicator-based allocation mechanisms and to a lesser extent in some countries using large-scale assessment exercises. However, there are at least three (probably more) potential solutions to this dilemma. The first of these is the minimax method. Minimax is a method in decision theory for minimizing the maximum possible loss (Clemen, 1996; Smith, 1988). Alternatively, it can be thought of as maximizing the minimum gain (maximin). It has also been extended to more complex games and to general decision making in the presence of uncertainty; that is, game theory (Nash, 1950). The second option is the wild card procedure. The wild card procedure is an approach to funding research proposals which is designed to reduce the shared bias problem in the evaluation of research proposals by sequestering a small amount of funds to ideas or proposals with large possible payoffs, even if the probability of success is likely to be small. The aim, therefore, is to “avoid the present tendency toward exclusion of heresy, especially in tight funding times, given the evidence, based on long experience, that people seen as heretics often lead the way to the big breakthroughs” (Scriven, 1991, p. 382). The third solution is to simply diminish, or eliminate entirely, the use of single indicator predictions of future performance (i.e., record of past performance) as an assurance of continued good performance (Scriven, 2007c).

Peers...Collective Wisdom or Shared Bias?

Federal agencies and private foundations in the USA and elsewhere rely on peer review to evaluate proposals for funding. Universities depend on it to make decisions about hiring, promotion, and tenure; they also use it to assess the quality of departments and programs and to make recommendations for eliminating or expanding departments. Journal and book editors use peer review to accept and reject manuscripts for publication. Most people in academia take for granted that decisions
about their work, their potential, and the course of their lives will be made by peer review (Eisenhart, 2002, p. 242).

The collective wisdom (sometimes, the shared bias) inherent in peer review has made the process the ‘gold standard’ for scholarly work, and peer review is likely to get more important as the task of separating signal from noise in the tidal wave of research becomes increasingly difficult. And, too often peer review is perceived as a mechanism, a set of practices and principles for allocating rewards and resources. It is, however, a severe oversimplification to reduce peer review to a measuring and allocation system. Despite its limitations (and its strengths), a partial list of the functions of peer review would include the following (Hackett, 1997):

1. A mechanism for improving research, both proposals and manuscripts
2. A forum for establishing priority claims and determining research priorities within research areas
3. A counterweight for the pursuit of originality and to ensure proper recognition of prior work
4. A procedure for allocating the scarce resources of research support, journal space, and recognition
5. A communication system that circulates research plans, increases the receptivity of others to forthcoming results, and provides reassurance and confidence to the proposing researcher
6. A point of entry for non-scientific considerations to influence science in a limited and controlled fashion
7. A quality control system for assuring non-expert users of published research that the work meets professional standards
8. An affirmation of professional authority and autonomy
9. An expression of the ideal of participation in knowledge production
10. A practice that affirms public trust in experts and experts’ trust in each other
In *Impure Science* (1992), Bell attempts to expose the overtly politicized nature of the United States’ peer review system. In one notable case, an archeologist was denied research support because a conspiracy of colleagues working on similar topics used the peer review system to unfairly thwart the archeologist’ grant application. In another, a research center was funded because a conspiracy of colleagues used the peer review system to unfairly promote its grant application. Essentially, Bell concluded that science is politicized and that peer review is often a conduit of bias.

While past scientific performance (indexed by publications and citations) is strongly correlated to future performance, peer ratings of proposals do not (Abrams, 1991). Since it is unlikely that proposal ratings are correlated with subsequent performance, Abrams (1991) argues that hurried evaluations by relatively poorly qualified judges should be abandoned and a system that awards research grants should include a moving average of peer ratings of prior work, citation and publication counts, with handicaps for older investigators, younger investigators, field-switchers, women, minorities, and geographic diversity. Hackett (1997), however, argues that burdening the peer review system with the task of fair treatment of young and old, black and white, men and women, and those located on the coasts and in the heartland will result in a system that “would appear arbitrary and illegitimate” (p. 53).

While Abram’s analysis of the shortcomings of the peer review system is generally convincing, he places unwarranted faith in publication and citation indicators as proxies of scientific quality. First, the practice of using publication and citation counts to evaluate the work of individual researchers is unstable; such indicators are more stable and valid only when applied to larger research collectives such as groups or disciplines, for instance. Second, peer judgments of papers are subject to many of the same distorting forces as peer review of research proposals; they are not always reliable and
valid and often influenced by author (and reviewer) characteristics. Whatever concreteness of judgment is gained by having a completed piece of work in hand is lost when the paper's results interact with referees' preferences and preconceptions.

Scientists and researchers are prone to disagree (especially in new or emerging fields of research). As Hackett (1997) explains, some researchers favor observational studies, other insist on experiments, some are clinicians, others rat runners, and they are only one sort of researcher and but one of many stakeholders in the peer review system. The fundamental problem is not to devise mechanisms for eliminating it (i.e., variability), but to make good use of this disagreement, not to treat it as embarrassing or aberrant. Too much energy is expended trying to homogenize peer ratings (not that this is undesirable, only sometimes not feasible), when, at their best, which is to say, when done by highly trained and highly experienced evaluators, they are rather good.

Finally, the concept of bias is seldom discussed, but interpreted in various ways (Langfeldt, 2002; Scriven, 1975). Some studies finding disagreements among peer reviewers interpret this as some sort of 'cognitive particularism' (Travis & Collins, 1991) or 'confirmatory bias' (Mahoney 1977), while others interpret disagreements as 'real and legitimate differences of opinion among experts about what good science is or should be' (Cole, Cole, & Simon, 1981). Such divergent interpretations reveal a lack of common understanding not only of the notion of bias, but also about what are legitimate considerations when evaluating research. Impartial assessments may be constructed (at least theoretically; see John Rawls' *A Theory of Justice*, 1971) and a prominent feature is the issue of anonymity. Briefly, the question is whether peer review is less biased when the authors' identity is withheld from reviewers and when the reviewers' identity is not revealed to authors and whether serious calibration can reduce undesirable variation in peer ratings.
The Holy Grail of Research Evaluation

It would be folly to fund research on the basis of impact metrics alone. Such measures are underdeveloped and are not robust surrogates for research impact. Some indicators may be used to support impact claims...But these are clustered among lower levels of impact such as engagement...industry funding, citations in government reports, creative works commissioned, and spin-off companies created...When it comes to demonstrating high levels of public value, some measures are proposed, like a reduction in regional crime rates. But how can a research group’s claim that its research caused this drop be validated? (Donovan, 2006, November 1).

The challenges implicit in identifying the true impacts or effects (e.g., social, economic, environmental) of most research have been addressed by generations of researchers and evaluators, with considerably variation in results (there are notable exceptions, particularly in the natural sciences; for example, research on vaccines which save countless lives). These challenges are intensified when evaluating so-called ‘non-science’ research such as the arts and humanities, as well as some disciplines in the social sciences. Nevertheless, the ability to identify these effects on different levels (e.g., macro, meso, micro) and from different perspectives in time (e.g., in the planning phase, in the developmental phase, in the dispositional phase, and in the historical phase) is not well developed. Measuring research impacts and effects is feasible, but the methods required are quite different from simply measuring inputs and outputs and, above all, is scarcely amenable to national summation. A major dilemma is that the detractors muddle the relationship between quality and impact. Quality is generally about academic consumption of research and is largely confined to scholarly publications. Impact is usually about end-user uptake and creating public value, a process that may begin or end with the publication of academic papers or books. Moreover, it is often about tailoring the communication (i.e., dissemination) of research outcomes to different audiences: an article in a practitioner journal, a report for a government agency, a creative
performance, a museum exhibition, contract research for industry, or a policy (Donovan, 2008).

**Concluding Remarks**

While there has been no other comparable work in this area, the findings regarding the national models should not be accepted as definitive. They do, however, provide a compelling and stark assessment of the variability in quality of different national models for evaluating government-financed research at a reasonable level of analysis; one that should provoke concern given the low ratings received by the majority of the models and even more so considering the stakes often involved (e.g., the course of lives sometimes swing in the balance). On the whole, however, this research has only scratched the surface of the numerous intricacies and complexities involved in national-level evaluations of researchers and their research. The methods in this dissertation also demonstrate a methodology of how such evaluations can be conducted.

However, as noted previously, in most nations there are serious flaws and weaknesses in their understanding and application of the reasoning and logic of evaluation. Evaluative reasoning consists of two distinct processes, which is close to the heart of scientific practice and the public costs and benefits of science. The first of these processes is evidentiary and the second inferential. The logic of evaluation, by comparison, can be reduced to matters such as its definition (i.e., systematically and objectively determining the merit, worth, or significance of things) and the definition of its major concepts (e.g., grading, ranking, scoring, apportioning, synthesis, barring, steeping, scaling, profiling), the nature of its relations to other subjects and other disciplines (i.e., the transdisciplinary model), and the rules of inference that govern it (e.g., beyond reasonable doubt; the balance or weight of evidence). Relatedly, there are
some ways in which the evaluation of research, notably the evaluation of proposals for
the funding of scientific research, can be much improved by using techniques from the
logic of evaluation. Therefore, it is clear that there are major ways in which common
processes in the evaluation and funding of all kinds of research can be improved with
large gains in money saved, good research done, and valuable results obtained.
APPENDIX A

Acronyms

This appendix is intended to serve as a point of reference for the many acronyms that occur throughout this dissertation. In the text of the dissertation the first appearance of a compound term is followed by its acronym in parentheses; for example, Organisation for Economic Co-Operation and Development (OECD). From that point forward only the acronym is used in text (i.e., OECD rather than Organisation for Economic Co-Operation and Development).

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>Central American Evaluation Society</td>
</tr>
<tr>
<td>AEA</td>
<td>American Evaluation Association</td>
</tr>
<tr>
<td>AERA</td>
<td>American Educational Research Association</td>
</tr>
<tr>
<td>AERES</td>
<td>Agency for the Evaluation of Research and Higher Education</td>
</tr>
<tr>
<td>AES</td>
<td>Australasian Evaluation Society</td>
</tr>
<tr>
<td>AiF</td>
<td>German Federation of Industrial Cooperative Research Associations</td>
</tr>
<tr>
<td>AIST</td>
<td>Advanced Industrial Science and Technology</td>
</tr>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>ANR</td>
<td>National Research Agency</td>
</tr>
<tr>
<td>APA</td>
<td>American Psychological Association</td>
</tr>
<tr>
<td>ARC</td>
<td>Academic and Research Commission/Concerted Research Actions</td>
</tr>
<tr>
<td>AS CR</td>
<td>Academy of Sciences of the Czech Republic</td>
</tr>
<tr>
<td>ATP</td>
<td>Advanced Technology Program</td>
</tr>
<tr>
<td>ATWS</td>
<td>average total weighted score</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>AU</td>
<td>Australia</td>
</tr>
<tr>
<td>AVČR</td>
<td>Academy of Sciences of the Czech Republic</td>
</tr>
<tr>
<td>BE</td>
<td>Belgium</td>
</tr>
<tr>
<td>BERD</td>
<td>GERD financed by business enterprise sector</td>
</tr>
<tr>
<td>BLK</td>
<td>Commission for Educational Planning and Research Promotion</td>
</tr>
<tr>
<td>BMBF</td>
<td>Bundesministerium für Bildung und Forschung</td>
</tr>
<tr>
<td>BOF</td>
<td>Special Research Fund</td>
</tr>
<tr>
<td>BRD</td>
<td>beyond reasonable doubt</td>
</tr>
<tr>
<td>BUGC</td>
<td>British University Grants Committee</td>
</tr>
<tr>
<td>CCG</td>
<td>Commonwealth Competitive Grants</td>
</tr>
<tr>
<td>CES</td>
<td>Canadian Evaluation Society</td>
</tr>
<tr>
<td>CI</td>
<td>composite index</td>
</tr>
<tr>
<td>CIP</td>
<td>Competitiveness and Innovation Framework Programme</td>
</tr>
<tr>
<td>CityU</td>
<td>City University of Hong Kong</td>
</tr>
<tr>
<td>CNE</td>
<td>National Committee for Evaluation</td>
</tr>
<tr>
<td>CNER</td>
<td>National Committee for Research Evaluation</td>
</tr>
<tr>
<td>CNRS</td>
<td>Center de la Recherche Scientifique</td>
</tr>
<tr>
<td>CORE</td>
<td>Center of Research Excellence</td>
</tr>
<tr>
<td>COSEUP</td>
<td>Committee on Science, Engineering, and Public Policy</td>
</tr>
<tr>
<td>CRE</td>
<td>Contribution to Research Environment</td>
</tr>
<tr>
<td>CSR</td>
<td>Committee for Scientific Research</td>
</tr>
<tr>
<td>CSTP</td>
<td>Council for Science and Technology Policy</td>
</tr>
<tr>
<td>CV</td>
<td>curriculum vita</td>
</tr>
<tr>
<td>CUHK</td>
<td>Chinese University of Hong Kong</td>
</tr>
<tr>
<td>CZ</td>
<td>Czech Republic</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>CZK</td>
<td>Czech Koruna</td>
</tr>
<tr>
<td>DBPR</td>
<td>double-blind peer review</td>
</tr>
<tr>
<td>DE</td>
<td>Germany</td>
</tr>
<tr>
<td>DeGEval</td>
<td>German Society for Evaluation</td>
</tr>
<tr>
<td>DENI</td>
<td>Department of Education for Northern Ireland</td>
</tr>
<tr>
<td>DEST</td>
<td>Department of Education, Science and Training</td>
</tr>
<tr>
<td>DETE</td>
<td>Department for Enterprise, Trade and Employment</td>
</tr>
<tr>
<td>DETYA</td>
<td>Department of Education, Training and Youth Affairs</td>
</tr>
<tr>
<td>DFA</td>
<td>discriminant function analysis</td>
</tr>
<tr>
<td>DFG</td>
<td>German Research Foundation</td>
</tr>
<tr>
<td>DG JRC</td>
<td>European Commission's Joint Research Center</td>
</tr>
<tr>
<td>DGRST</td>
<td>General Directorate for Scientific and Technical Research</td>
</tr>
<tr>
<td>DGTRE</td>
<td>Directorate General for Technologies, Research and Energy</td>
</tr>
<tr>
<td>DHHS</td>
<td>Department of Health and Human Services</td>
</tr>
<tr>
<td>DOD</td>
<td>Department of Defense</td>
</tr>
<tr>
<td>DOE</td>
<td>Department of Energy</td>
</tr>
<tr>
<td>DRME</td>
<td>Directorate for Research and Test Facilities</td>
</tr>
<tr>
<td>EAG</td>
<td>expert advisory group</td>
</tr>
<tr>
<td>EC</td>
<td>The Evaluation Center</td>
</tr>
<tr>
<td>EES</td>
<td>European Evaluation Society</td>
</tr>
<tr>
<td>EOCP</td>
<td>Economic Competitiveness Operative Program</td>
</tr>
<tr>
<td>EvC</td>
<td>Evaluation Committee</td>
</tr>
<tr>
<td>HER</td>
<td>Education and Human Resources Directorate</td>
</tr>
<tr>
<td>EC</td>
<td>The Evaluation Center</td>
</tr>
<tr>
<td>ECOP</td>
<td>Economic Competitiveness Operative Program</td>
</tr>
</tbody>
</table>
EEA  evaluation of educational activities by academic field
EFTS. equivalent-fulltime students
Enet Evaluation Network
EP evidence portfolio
EPO European Patent Office
ERA European Research Area; evaluation of research by academic field
ERI external research income
ERS Evaluation Research Society
ETS Educational Testing Services
EU European Union
FAS Swedish Council for Working Life and Social Research
FDA Food and Drug Administration
FHEEC Finland Higher Education Evaluation Council
FhG Fraunhofer Gesellschaft
FI Finland
FINHEEC Finland Higher Education Evaluation Council
FNRS National Fund for Scientific Research
FORMAS Council for Environment, Agricultural Sciences and Spatial Planning
FP7 Seventh Framework Programme
FR France
FRG Federal Republic of Germany
FRST Foundation for Research, Science and Technology
FTE full-time equivalent
FWO Fonds Wetenschappelijk Onderzoek
GAO General Accounting Office
government support for R&D that has been developed using budget data

GDP: gross domestic product

GDR: German Democratic Republic

GERD: gross domestic expenditure on R&D

GOE: grade of execution

GPRA: Government Performance and Results Act of 1993

HEA: Higher Education Authority

HEFC: Higher Education Funding Council

HEFCE: Higher Education Council for England

HEFCW: Higher Education Funding Council for Wales

HEI: Higher Education Institution

HERO: Higher Education & Research Opportunities for the United Kingdom

HGF: Helmholtz-Gemeinschaft Deutscher Forschungszentren

HK: Hong Kong

HKBU: Hong Kong Baptist University

HKIEd: Hong Kong Institute of Education

HKU: University of Hong Kong

HKUST: Hong Kong University of Science & Technology

HOBEK: Hoger Onderwijs Bekostigingsmodel

HR: human resources

HRB: Health Research Board

HRC: Health Research Council

HSIRB: Human Subjects Institutional Review Board

HU: Hungary

IAI: Independent Administrative Institutions
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSTI</td>
<td>Irish Council for Science, Technology and Innovation</td>
</tr>
<tr>
<td>ICT</td>
<td>Information Communications Technology</td>
</tr>
<tr>
<td>IDA</td>
<td>Industrial Development Agency</td>
</tr>
<tr>
<td>IDCSTI</td>
<td>Inter Departmental Committee on Science, Technology and Innovation</td>
</tr>
<tr>
<td>IDPE</td>
<td>Interdisciplinary Ph.D. in Evaluation</td>
</tr>
<tr>
<td>IE</td>
<td>Ireland</td>
</tr>
<tr>
<td>IES</td>
<td>Institute for Educational Science</td>
</tr>
<tr>
<td>IFQ</td>
<td>Institute for Research Information and Quality Assurance</td>
</tr>
<tr>
<td>IGS</td>
<td>Institutional Grants Scheme</td>
</tr>
<tr>
<td>IIHS</td>
<td>Insurance Institute for Highway Safety</td>
</tr>
<tr>
<td>IMEC</td>
<td>Interuniversity MicroElectronics Center</td>
</tr>
<tr>
<td>IO</td>
<td>industrial and organizational psychology</td>
</tr>
<tr>
<td>IOM</td>
<td>Institute of Medicine</td>
</tr>
<tr>
<td>IPC</td>
<td>Innovation Policy Council</td>
</tr>
<tr>
<td>IRCHHS</td>
<td>Irish Research Council for the Humanities and the Social Sciences</td>
</tr>
<tr>
<td>IRCSET</td>
<td>Irish Research Council for Science, Engineering and Technology</td>
</tr>
<tr>
<td>IRSIB</td>
<td>Institute for the Encouragement of Scientific Research and Innovation</td>
</tr>
<tr>
<td>ISI</td>
<td>Institute for Scientific Information</td>
</tr>
<tr>
<td>ITPS</td>
<td>Institute for Growth Policy Studies</td>
</tr>
<tr>
<td>IUQB</td>
<td>Irish Universities Quality Board</td>
</tr>
<tr>
<td>IWT</td>
<td>Institute for the Promotion of Innovation by Science and Technology</td>
</tr>
<tr>
<td>JBR</td>
<td>Research and Development Units</td>
</tr>
<tr>
<td>JP</td>
<td>Japan</td>
</tr>
<tr>
<td>JPO</td>
<td>Japan Patent Office</td>
</tr>
<tr>
<td>JREN</td>
<td>Japan Research Evaluation Network</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Name</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>JSPS</td>
<td>Japan Society for the Promotion of Science</td>
</tr>
<tr>
<td>JST</td>
<td>Japan Science and Technology Agency</td>
</tr>
<tr>
<td>KBN</td>
<td>State Committee for Scientific Research</td>
</tr>
<tr>
<td>KEC</td>
<td>Key Evaluation Checklist</td>
</tr>
<tr>
<td>KKS</td>
<td>Knowledge Foundation</td>
</tr>
<tr>
<td>KNAW</td>
<td>Royal Netherlands Academy of Arts and Sciences</td>
</tr>
<tr>
<td>KPR</td>
<td>National Framework Program</td>
</tr>
<tr>
<td>LPRD</td>
<td>Long-Term Principal Research Directions</td>
</tr>
<tr>
<td>LU</td>
<td>Lingnan University</td>
</tr>
<tr>
<td>M&amp;A</td>
<td>monitoring and analysis</td>
</tr>
<tr>
<td>MBO</td>
<td>management by objective</td>
</tr>
<tr>
<td>MEP</td>
<td>Manufacturing Extension Partnership</td>
</tr>
<tr>
<td>MESSC</td>
<td>Ministry of Education, Science, Sports and Culture</td>
</tr>
<tr>
<td>METI</td>
<td>Ministry of Economy, Trade and Industry</td>
</tr>
<tr>
<td>MEXT</td>
<td>Ministry of Education, Culture, Sports, Science and Technology</td>
</tr>
<tr>
<td>MK</td>
<td>Finnish Markka</td>
</tr>
<tr>
<td>MNiE</td>
<td>Ministry of Science and Education</td>
</tr>
<tr>
<td>MnIL</td>
<td>Ministry of Science and Information Technologies</td>
</tr>
<tr>
<td>MniSW</td>
<td>Ministry of Science and Higher Education</td>
</tr>
<tr>
<td>MPG</td>
<td>Max Planck Gesellschaft</td>
</tr>
<tr>
<td>MTA</td>
<td>Hungarian Academy of Sciences</td>
</tr>
<tr>
<td>NAE</td>
<td>National Academy of Engineering</td>
</tr>
<tr>
<td>NAS</td>
<td>National Academy of Sciences</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
</tr>
<tr>
<td>NATO</td>
<td>North Atlantic Treaty Organization</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>NDP</td>
<td>National Development Plan</td>
</tr>
<tr>
<td>NEDO</td>
<td>New Energy and Industrial Technology Development Organisation</td>
</tr>
<tr>
<td>NFR</td>
<td>Swedish Natural Science Research Council</td>
</tr>
<tr>
<td>NGO</td>
<td>non-governmental organization</td>
</tr>
<tr>
<td>NHMRC</td>
<td>National Health and Medical Research Council</td>
</tr>
<tr>
<td>NHST</td>
<td>null hypothesis significance testing</td>
</tr>
<tr>
<td>NIAD</td>
<td>National Institute for Academic Degrees and University Evaluation</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NIS</td>
<td>newly independent state</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute of Standards and Technology</td>
</tr>
<tr>
<td>NISTEP</td>
<td>National Institute of Science and Technology Policy</td>
</tr>
<tr>
<td>NL</td>
<td>Netherlands</td>
</tr>
<tr>
<td>NPM</td>
<td>New Public Management</td>
</tr>
<tr>
<td>NRC</td>
<td>National Research Council</td>
</tr>
<tr>
<td>NRDP</td>
<td>National Research and Development Policy</td>
</tr>
<tr>
<td>NRP</td>
<td>National Research Program</td>
</tr>
<tr>
<td>NSF</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>NSTC</td>
<td>National Science and Technology Council</td>
</tr>
<tr>
<td>NOW</td>
<td>Netherlands Organisation for Scientific Research</td>
</tr>
<tr>
<td>NRP</td>
<td>National Research Program</td>
</tr>
<tr>
<td>NWS</td>
<td>numeric weight and sum</td>
</tr>
<tr>
<td>NZ</td>
<td>New Zealand</td>
</tr>
<tr>
<td>NZD</td>
<td>New Zealand Dollar</td>
</tr>
<tr>
<td>OE</td>
<td>Overseas Experience</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-Operation and Development</td>
</tr>
</tbody>
</table>
OLS ordinary least squares
OMB Office of Management and Budget
ONR Office of Naval Research
OST Office of Science and Technology
OSTP Office of Science and Technology Policy
PART Program Assessment Rating Tool
PBI past behavioral interview
PBRF Performance-Based Research Fund
PCAST Presidents Council of Advisors on Science and Technology
PC Programme Committee
PCFC Polytechnics and Colleges Funding Council
PE peer esteem
PEMD Program Evaluation and Methodology Division
PL Poland
PLN Polish Zlotych
PolyU Hong Kong Polytechnic University
PPBS Planning-Programming-Budgeting System
PRES Research and Higher Education Clusters
PRO Public Research Organisation
PRTLI Program for Research in Third Level Institutes
QAQI Quality Assurance/Quality Improvement
QE Quality Evaluation
QWS qualitative weight and sum
R Research
R&D Research and Development
RAE  Research Assessment Exercise
RAG  research ad hoc group
RAWB Dutch Advisory Council for Science Policy
RBTE Research-Based Teacher Evaluation
RCT  randomized controlled trial
RDC  postgraduate research degree completions
REPP Research Evaluation and Policy Project
ResEval World Research Evaluation Network
RFM  relative funding model
RFP  request for proposals
RGC  Research Grants Council
RIKEN Institute of Physical and Chemical Research
RO  research output
RPE  performance measurement and evaluation
RQ  research quantum
RPC  Research Policy Council
RQF  Research Quality Framework
RS  raw score
RTD  Research, Technology, and Development
RTDI Research, Technology, Development, and Innovation
RTRA Thematic Advanced Research Networks
RTS  research training scheme
S&T  science and technology
SBPR single-blind peer review
SCI  Science Citation Index
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDA</td>
<td>Self Defense Agency</td>
</tr>
<tr>
<td>SE</td>
<td>Sweden</td>
</tr>
<tr>
<td>SEK</td>
<td>Swedish Krona</td>
</tr>
<tr>
<td>SF</td>
<td>Structural Fund</td>
</tr>
<tr>
<td>SFI</td>
<td>Science Foundation Ireland</td>
</tr>
<tr>
<td>SHEFC</td>
<td>Scottish Higher Education Funding Council</td>
</tr>
<tr>
<td>SME</td>
<td>small and medium-sized enterprises</td>
</tr>
<tr>
<td>SNSB</td>
<td>Swedish National Space Board</td>
</tr>
<tr>
<td>SOO</td>
<td>Steunpunt O&amp;O Statistieken</td>
</tr>
<tr>
<td>SOUMU</td>
<td>Ministry for Internal Affairs and Communications</td>
</tr>
<tr>
<td>SSF</td>
<td>Swedish Foundation for Strategic Research</td>
</tr>
<tr>
<td>ST&amp;I</td>
<td>science, technology, and innovation</td>
</tr>
<tr>
<td>STA</td>
<td>Science and Technology Agency</td>
</tr>
<tr>
<td>STABEK</td>
<td>Stabiele Bekostiging</td>
</tr>
<tr>
<td>STS</td>
<td>State Technical Services Program</td>
</tr>
<tr>
<td>T</td>
<td>Teaching</td>
</tr>
<tr>
<td>TEAC</td>
<td>Tertiary Education Advisory Commission</td>
</tr>
<tr>
<td>TEO</td>
<td>Tertiary Education Organisations</td>
</tr>
<tr>
<td>TEP</td>
<td>Technology Foresight Program</td>
</tr>
<tr>
<td>TIG</td>
<td>Topical Interest Group</td>
</tr>
<tr>
<td>TOR</td>
<td>terms of reference</td>
</tr>
<tr>
<td>TQM</td>
<td>total quality management</td>
</tr>
<tr>
<td>TWS</td>
<td>total weighted score</td>
</tr>
<tr>
<td>UA</td>
<td>University of Auckland</td>
</tr>
<tr>
<td>UFC</td>
<td>Universities Funding Council</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>UGC</td>
<td>University Grants Committee</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UNS</td>
<td>Unified National System</td>
</tr>
<tr>
<td>UoA</td>
<td>unit of assessment</td>
</tr>
<tr>
<td>UPGC</td>
<td>University and Polytechnic Grants Committee</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USD</td>
<td>United States Dollar</td>
</tr>
<tr>
<td>USDOE</td>
<td>United States Department of Education</td>
</tr>
<tr>
<td>USPTO</td>
<td>United States Patent and Trademark Office</td>
</tr>
<tr>
<td>UwTE</td>
<td>university-wide thematic evaluation</td>
</tr>
<tr>
<td>VINNOVA</td>
<td>Swedish Agency for Innovation Systems</td>
</tr>
<tr>
<td>VIS</td>
<td>Flemish Cooperative Innovation Networks</td>
</tr>
<tr>
<td>VNU</td>
<td>Verbund Norddeutscher Universitäten</td>
</tr>
<tr>
<td>VR</td>
<td>Swedish Research Council</td>
</tr>
<tr>
<td>VSNNU</td>
<td>Association of the Netherlands Universities</td>
</tr>
<tr>
<td>WGL</td>
<td>Wissenschaftsgemeinschaft Wilhelm-Gottfried-Leibniz</td>
</tr>
<tr>
<td>WMU</td>
<td>Western Michigan University</td>
</tr>
<tr>
<td>WoS</td>
<td>Web-of-Science</td>
</tr>
<tr>
<td>WR</td>
<td>Science Council</td>
</tr>
<tr>
<td>WREN</td>
<td>Washington Research Evaluation Network</td>
</tr>
<tr>
<td>WS</td>
<td>weighted score</td>
</tr>
<tr>
<td>WWC</td>
<td>What Works Clearinghouse</td>
</tr>
<tr>
<td>ZBB</td>
<td>zero-based budgeting</td>
</tr>
</tbody>
</table>
APPENDIX B

Judge Socio-Demographic Questionnaire

Please select only one response per item or write your answer in the space provided.

1. What is your age? ______

2. What is your gender?
   - Male
   - Female

3. Which of the following best describes your race/ethnicity?
   - African American/Black
   - Asian
   - American Indian/Alaska Native
   - Caucasian/White
   - East Asian/Middle Eastern
   - Hispanic/Latino
   - Māori
   - Native Hawaiian
   - Pasifika
   - Other (please indicate): ____________________

4. What is your country of birth? ____________________

5. Highest academic degree attained to date:
   - Doctoral degree
   - Master’s degree
   - Bachelor’s degree
   - Other (please indicate): ____________________
6. In general, which of the following best characterizes the extent of your professional evaluation experience, if any?
   - None
   - Very little
   - Some
   - A great deal
   - Extensive

7. In general, which of the following best characterizes the extent of your professional research experience, if any?
   - None
   - Very little
   - Some
   - A great deal
   - Extensive

8. What is your cognate or disciplinary area of expertise? (please describe)

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

9. What other experience do you have that is relevant to this study? (please describe)

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________
APPENDIX C

Judge Scoring Sheet

For each of the following items please score/rate the country’s evaluation model/system from 0-10, where 0 = absence of merit and 10 = excellent. Please do not score/rate any item more than once. Please do not skip any items. Use your best judgment on the information provided. Finally, you will find a “Narrative Critique” section at the end of this scoring sheet; please read the instructions carefully and complete the section as described.

Country Code:
Judge:

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Score/Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Takes appropriate factors into account</td>
<td></td>
</tr>
<tr>
<td>2. Weights factors in a transparent, defensible, appropriate manner</td>
<td></td>
</tr>
<tr>
<td>3. Does not confuse grading, ranking, scoring, and apportioning</td>
<td></td>
</tr>
<tr>
<td>4. Conclusions are logically and demonstrably correct, justifiable</td>
<td></td>
</tr>
<tr>
<td>5. Is capable of replication or has the quality of being repeatable</td>
<td></td>
</tr>
<tr>
<td>6. Is transparent</td>
<td></td>
</tr>
<tr>
<td>7. Is impartial and unbiased</td>
<td></td>
</tr>
<tr>
<td>8. Has defensible accounts for lack of conflict of interest</td>
<td></td>
</tr>
<tr>
<td>9. Is conducted by those with adequate expertise</td>
<td></td>
</tr>
<tr>
<td>10. Is externally credible, believable to right-to-know audiences</td>
<td></td>
</tr>
<tr>
<td>11. Is relevant, fit for purpose</td>
<td></td>
</tr>
<tr>
<td>12. Is timely</td>
<td></td>
</tr>
<tr>
<td>13. Is easy to apply</td>
<td></td>
</tr>
<tr>
<td>14. Is easy to understand</td>
<td></td>
</tr>
<tr>
<td>15. Provides feedback to those evaluated</td>
<td></td>
</tr>
<tr>
<td>16. Money costs are reasonable given the benefits produced</td>
<td></td>
</tr>
</tbody>
</table>
17. Costs are reasonable in terms of time
18. Costs are reasonable in terms of specialist expertise
19. Costs are reasonable for those submitting information
20. Payoff is substantial
21. Able to detect or uncover fraud or misconduct
22. Allows an appeal process
23. Deals fairly with new as well as experienced researchers
24. Gives a complete and fair assessment
25. Is independently metaevaluated
26. Overall best judgment
27. What country do you think this is?

Narrative Critique: This section should reflect your overall assessment. It should summarize the salient features and the primary reasons for your scores/ratings. This critique should highlight the strengths and weaknesses of the research evaluation model/system. If you need additional space, please continue on the back of this page.

1. Please describe the most salient features and the primary reasons for your scores/ratings.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

2. What were the key strengths and weaknesses of the model/system?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
3. What was missing from the model/system?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

4. What would you suggest or recommend for improving the model/system?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
APPENDIX D

Judge Consent Form Version A

Western Michigan University
Department: The Evaluation Center
Principal Investigator: Dr. Michael Scriven
Student Investigator: Chris L. S. Coryn

You have been invited to participate in a research project entitled “A Comparative Study of International Research Evaluation Models.” This research is intended to rate, profile, and rank the quality of research evaluation models currently used in sixteen countries. This project is part of Chris L. S. Coryn’s dissertation project.

You will be asked to attend a full-day session with Chris L. S. Coryn. In all, you will be asked for approximately eight hours of your time. Prior to attending the session you will also be asked to complete a brief sociodemographic questionnaire. The sociodemographic questionnaire will ask you to provide general information about yourself, such as your age, level of education, and research and/or evaluation experience. You will also be asked to submit your current curriculum vita or résumé prior to the session.
You will be asked to meet Chris L. S. Coryn at The Evaluation Center. The first part of the session will involve an overview of the study and training for the tasks that you will be asked to complete. During this part of the session you will be asked to rate two or three countries’ research evaluation models. The first part of the session will begin at 8:30 AM and end at 12:30 PM. A two-hour lunch break will take place from 12:30 PM to 2:30 PM. Lunch will be provided for you. Following the lunch break, the second part of the session will take place. During the second part of the session you will be asked to participate in a group activity and to openly discuss the activity with others in the session. The second part of the session will begin at 2:30 PM and end at 6:30 PM. In both sessions you will be asked to read narratives describing the research evaluation models of two or three countries. These narratives range in length from five to ten pages. You will also be asked to complete a rating form for each of the countries assigned to you.

There are no known risks associated with participating in this study. However, there are inconveniences related to the time required to participate.

One way in which you may benefit from this activity is having the chance to participate in a ‘real-world’ evaluation as well as influencing international research evaluation policy.

All of the information collected from you is confidential. That means that your name will not appear on any papers on which this information is recorded. The forms will all be coded, and Chris L. S. Coryn will keep a separate master list with the names of participants and the corresponding code numbers. Once the data are collected and analyzed, the master list will be destroyed. All other forms will be retained for at least three years in a locked file in the principal investigator’s office.
You may refuse to participate or quit at any time during the study without prejudice or penalty. If you have any questions or concerns about this study, you may contact either Chris L. S. Coryn at 269-387-5920 or Dr. Michael Scriven at 269-387-5906. You may also contact the chair of Human Subjects Institutional Review Board at 269-387-8293 or the vice president for research at 269-387-8298 with any concerns that you have.

This consent document has been approved for use for one year by the Human Subjects Institutional Review Board as indicated by the stamped date and signature of the board chair in the upper right corner. Do not participate in this study if the stamped date is more than one year old.

You will get two copies of the consent document—one to sign and return and one to keep. Your signature below indicates that you have read and/or had explained to you the purpose and requirements of the study and that you agree to participate.

_________________________  ____________
Signature                  Date

Consent obtained by:  ____________  ____________
initials of researcher  Date
You have been invited to participate in a research project entitled "A Comparative Study of International Research Evaluation Models." This research is intended to rate, profile, and rank the quality of research evaluation models currently used in sixteen countries. This project is part of Chris L. S. Coryn’s dissertation project.

You will be asked to attend a full-day session with Chris L. S. Coryn. In all, you will be asked for approximately eight hours of your time. Prior to attending the session you will also be asked to complete a brief sociodemographic questionnaire. The sociodemographic questionnaire will ask you to provide general information about yourself, such as your age, level of education, and research and/or evaluation experience. You will also be asked to submit your current curriculum vita or résumé prior to the session.
You will be asked to meet Chris L. S. Coryn at The University of Auckland, Faculty of Education. The first part of the session will involve an overview of the study and training for the tasks that you will be asked to complete. During this part of the session you will be asked to rate two or three countries' research evaluation models. The first part of the session will begin at 8:30 AM and end at 12:30 PM. A two-hour lunch break will take place from 12:30 PM to 2:30 PM. Lunch will be provided for you. Following the lunch break, the second part of the session will take place. During the second part of the session you will be asked to participate in a group activity and to openly discuss the activity with others in the session. The second part of the session will begin at 2:30 PM and end at 6:30 PM. In both sessions you will be asked to read narratives describing the research evaluation models of two or three countries. These narratives range in length from five to ten pages. You will also be asked to complete a rating form for each of the countries assigned to you.

There are no known risks associated with participating in this study. However, there are inconveniences related to the time required to participate.

One way in which you may benefit from this activity is having the chance to participate in a 'real-world' evaluation as well as influencing international research evaluation policy.

All of the information collected from you is confidential. That means that your name will not appear on any papers on which this information is recorded. The forms will all be coded, and Chris L. S. Coryn will keep a separate master list with the names of participants and the corresponding code numbers. Once the data are collected and analyzed, the master list will be destroyed. All other forms will be retained for at least three years in a locked file in the principal investigator's office.
You may refuse to participate or quit at any time during the study without prejudice or penalty. If you have any questions or concerns about this study, you may contact either Chris L. S. Coryn at 269-387-5920 or Dr. Michael Scriven at 269-387-5906. You may also contact the chair of Human Subjects Institutional Review Board at 269-387-8293 or the vice president for research at 269-387-8298 with any concerns that you have.

This consent document has been approved for use for one year by the Human Subjects Institutional Review Board as indicated by the stamped date and signature of the board chair in the upper right corner. Do not participate in this study if the stamped date is more than one year old.

You will get two copies of the consent document—one to sign and return and one to keep. Your signature below indicates that you have read and/or had explained to you the purpose and requirements of the study and that you agree to participate.

_________________________        ____________
Signature                      Date

Consent obtained by:           ____________        ____________
    initials of researcher    Date
Date: January 29, 2007

To: Michael Scriven, Principal Investigator
    Chris Coryn, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number: 07-01-05

This letter will serve as confirmation that your research project entitled "" has been approved under the expedited category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note that you may only conduct this research exactly in the form it was approved. You must seek specific board approval for any changes in this project. You must also seek reapproval if the project extends beyond the termination date noted below. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: January 29, 2008
REFERENCES


Commonwealth of Australia. (2005f). RQAF team support paper: Discussion paper on the rating scale to be used in the RQF. Canberra, Australia: Department of Education, Science and Training.


Zilhany, P., & Láng, I. (1997). The evaluation of research institutes in Hungary. In M. S. Frankel & J. Cave (Eds.), *Evaluating science and scientists: An East-West dialogue on*
research evaluation in post-Communist Europe (pp. 82-92). Budapest, Hungary: Central European University Press.

