This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.
Measuring research
A guide to research evaluation frameworks and tools

Susan Guthrie, Watu Wamae, Stephanie Diepeveen,
Steven Wooding and Jonathan Grant

RAND EUROPE
Measuring research

A guide to research evaluation frameworks and tools

Susan Guthrie, Watu Wamae, Stephanie Diepeveen, Steven Wooding and Jonathan Grant

Prepared for the Association of American Medical Colleges

MG-1217-AAMC
July 2013
The research described in this report was prepared for the Association of American Medical Colleges.
The Association of American Medical Colleges (AAMC) asked RAND Europe to conduct a detailed synthesis of existing and previously proposed frameworks and indicators used to evaluate research. The purpose of this review is to stimulate debate among US policymakers and the leadership of US medical schools and teaching hospitals on how research outcomes can be measured in different contexts and to address different needs and ultimately account to US taxpayers and others for the returns on their investment in research. The review also aims to provide a guide to the key considerations that need to be taken into account when developing approaches to research evaluation, to be used as a resource by research funders, managers and policymakers.

This report outlines the trade-offs that have to be taken into account and the contextual factors that need to be understood, drawing on experience of international approaches to research evaluation. In addition, a detailed overview of six research evaluation frameworks of particular interest and relevance to AAMC is provided, along with a brief overview of a further eight frameworks, and discussion of the main tools used in research evaluation. The report is likely to be of interest to policymakers, research funders, institutional leaders and research managers.

RAND Europe is a not-for-profit public policy research institute that aims to improve policy and decisionmaking through research and analysis. RAND Europe’s clients include European governments, institutions, NGOs and firms with a need for rigorous, independent, multi-disciplinary analysis. This report has been peer-reviewed in accordance with RAND’s quality assurance standards.

For more information about RAND Europe or this document, please contact:

Dr Susan Guthrie
RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
United Kingdom
Tel. +44 (1223) 353 329
sguthrie@rand.org
Contents

Preface .......................................................................................................................... iii
Table of Figures .......................................................................................................... vii
Table of Tables .......................................................................................................... viii
Executive summary .................................................................................................... ix
Glossary ....................................................................................................................... xii

CHAPTER 1
Introduction .................................................................................................................. 1
Structure and content of this report .......................................................................... 1
Frameworks and tools reviewed ................................................................................. 2

CHAPTER 2
Challenges and trade-offs in research evaluation ..................................................... 5
Characteristics of research evaluation approaches .................................................... 5
  Analysis of framework characteristics .................................................................. 8
  Analysis of tool characteristics ............................................................................ 9
Stage of measurement .................................................................................................. 12
Approaches to common challenges .......................................................................... 13
Credibility and acceptability ...................................................................................... 14
Challenges of implementation ................................................................................... 16
The future of research evaluation ............................................................................ 16

CHAPTER 3
Building a new research evaluation framework ....................................................... 19
Purpose ....................................................................................................................... 20
Characteristics of the framework .............................................................................. 20
Addressing the pitfalls ............................................................................................... 21
Consider the context .................................................................................................. 21
Select the tools .......................................................................................................... 22
Level of aggregation .................................................................................................. 22
Implementation .......................................................................................................... 23
  Summary: example case ....................................................................................... 23
  Decision tree ........................................................................................................ 24
Table of figures

**Figure 2.1** Comparing characteristics and purposes of research evaluation frameworks .................................................. 9  
**Figure 2.2** Mapping of tools against Group 1 and Group 2 characteristics ................................................................. 10  
**Figure 2.3** Unit of aggregation used in six frameworks investigated in detail .................................................................. 12  
**Figure 2.4** Measures used in six research evaluation frameworks .................................................................................. 13  
**Figure 3.1** A decision tree for developing a research evaluation framework .............................................................. 25  
**Figure A1** Scatter plot matrix of the correlations between the characteristics of frameworks .................................... 45  
**Figure A2** Scatter plot to determine the number of factors used in the analysis of the characteristics of frameworks ............ 47  
**Figure A3** Plot showing the alignment of the six frameworks investigated in detail with the five most significant factors ........................................................................................................................................... 49  
**Figure A4** CAHS Logic Model Framework built on a payback logic model ................................................................. 54  
**Figure A5** The CAHS framework .................................................................................................................................. 58  
**Figure A6** Impact categories cross mapped for the biomedical research pillar to create a specialised framework .......... 59  
**Figure A7** A SWOT analysis for the CAHS framework .................................................................................................. 61  
**Figure A8** Example of aggregated output from ERA 2010: data and rating summary ......................................................... 66  
**Figure A9** Example of aggregated output from ERA 2010: contribution to national landscape (%) ................................................. 67  
**Figure A10** A SWOT analysis for ERA ............................................................................................................................. 69  
**Figure A11** Template for collecting financial data for the NIHR Dashboard framework ................................................. 72  
**Figure A12** The integrated NIHR Dashboard structure .................................................................................................... 73  
**Figure A13** Conceptual approach for developing a performance monitoring framework .............................................. 74  
**Figure A14** A SWOT analysis for the NIHR Dashboard ................................................................................................... 76  
**Figure A15** A SWOT analysis for REF .......................................................................................................................... 87  
**Figure A16** A SWOT analysis for Productive Interactions ................................................................................................ 96  
**Figure A17** How STAR METRICS collects data from research institutions during Level 1 ............................................. 99  
**Figure A18** Example of data visualisation for an individual institution ........................................................................... 101  
**Figure A19** Example of data visualisation for aggregated data .......................................................................................... 102  
**Figure A20** A SWOT analysis for STAR METRICS ........................................................................................................ 104  
**Figure A21** A basic logic model .......................................................................................................................................... 150  
**Figure A22** The Buxton-Hanney Payback Framework .................................................................................................... 152  
**Figure A23** Types of survey .............................................................................................................................................. 163  
**Figure A24** An impact array for a research portfolio ....................................................................................................... 167  
**Figure A25** Classification of information visualisation techniques ...................................................................................... 170
Table of tables

Table 2.1 Purposes of the six frameworks investigated in detail ....................................................... 7
Table 2.2 The desirable characteristics of advocacy, analysis, accountability and allocation ............. 7
Table 2.3 Cross mapping of tools used to six evaluation frameworks ................................................. 11
Table A1 Summary table of the frameworks reviewed in this study .................................................. 37
Table A2 Summary table of the tools reviewed in this study ............................................................. 39
Table A3 Scoring of six frameworks against the characteristics identified ......................................... 42
Table A4 Scoring of all frameworks included in the analysis against characteristics identified ........ 44
Table A5 The level of correlation between characteristics of research evaluation frameworks .......... 46
Table A6 Loadings of the characteristics against each of the factors .................................................. 47
Table A7 Scoring of research evaluation tools against characteristics ............................................... 49
Table A8 Correlations between characteristics of research evaluation tools ..................................... 50
Table A9 Preferred indicators and metrics for impact categories of the CAHS framework ............... 57
Table A10 Time periods for different sets of indicators in ERA ....................................................... 65
Table A11 ERA rating scale 2010 ..................................................................................................... 65
Table A12 NIHR implementation plans .............................................................................................. 71
Table A13 Fifteen indicators of the NIHR pilot Dashboard ................................................................. 75
Table A14 Number of case studies required in submissions to the REF ........................................... 81
Table A15 Timeline for deployment of the REF ................................................................................ 84
Table A16 Case studies conducted as part of the SIAMPI study, by country and subject area .......... 90
Table A17 Examples of Productive Interactions, and their relationship to impact and stakeholders ........................................................................................................ 92
Table A18 Data elements requested at Level 1 of STAR METRICS .................................................. 100
Table A19 Assessment of the current state of the AERES research unit ........................................... 106
Table A20 Prospects and opportunities for developing research excellence of the AERES research unit .............................................................................................................................. 106
Table A21 CDMRP taxonomy system of research outcomes ............................................................. 109
Table A22 Overview of the CDMRP coding system .......................................................................... 109
Table A23 Rating levels of the NRF .................................................................................................. 115
Table A24 Assessment criteria, sub-criteria and aspects to be considered for the SEP ..................... 121
Executive summary

Interest in and demand for the evaluation of research is increasing internationally. Several factors account for this: agendas for accountability and good governance and management, and fiscal austerity in many countries. There is a need to show that policymaking is evidence based and, in the current economic climate, to demonstrate accountability for the investment of public funds in research. This is complemented by a shift in the type of evaluation needed: from the traditional, summative, assessment to more formative evaluations and those covering wider outputs from research.

Given this growing need for effective and appropriate evaluation, it is increasingly important to understand how research can and should be evaluated in different contexts and to meet different needs. The purpose of this study was to understand the challenges and trade-offs in evaluating research, by looking at examples of frameworks and tools for evaluating research used internationally. The frameworks and tools we explored in detail are summarised in Annex A. Our key findings are as follows.

Designing a research evaluation framework requires trade-offs: there is no silver bullet.

By investigating some of the characteristics of research evaluation frameworks, and the relationships between them, we have identified some trade-offs in developing a framework to evaluate research:

- Quantitative approaches (those which produce numerical outputs) tend to produce longitudinal data, do not require judgement or interpretation and are relatively transparent, but they have a high initial burden (significant work may be required at the outset to develop and implement the approach).
- Formative approaches (those focusing on learning and improvement rather than assessing the current status) tend to be comprehensive, evaluating across a range of areas, and flexible, but they do not produce comparisons between institutions.
- Approaches that have a high central burden (requiring significant work on the part of the body organising the evaluation process) tend not to be suitable for frequent use.
- Approaches that have been more fully implemented tend to have a high level of central ownership (by either the body organising the evaluation, or some other body providing oversight of the process).
- Frameworks that place a high burden on participants require those participants to have a high level of expertise (or should provide capacity building and training to achieve this).

To be effective, the design of the framework should depend on the purpose of the evaluation: advocacy, accountability, analysis and/or allocation.

Research evaluation aims to do one, or more, of the following:

- advocate: to demonstrate the benefits of supporting research, enhance understanding of research and its processes among policymakers and the public, and make the case for policy and practice change
- show accountability: to show that money and other resources have been used efficiently and effectively, and to hold researchers to account
• analyse: to understand how and why research is effective and how it can be better supported, feeding into research strategy and decision-making by providing a stronger evidence base
• allocate: to determine where best to allocate funds in the future, making the best use possible of a limited funding pot.

The aim(s) of the evaluation process should be decided at the outset, as this will influence most other decisions in the development of the research evaluation approach. For example, our analysis suggests that frameworks that are used to inform allocation often require comparisons to be made between institutions, and are therefore summative rather than formative, and are unlikely to be comprehensive. By contrast, analysis typically requires a more formative approach in order to show ‘how’ and ‘why’ the research is effective, rather than just provide a summative measure of performance. This suggests that frameworks for allocation are not likely to be well suited to use for analysis, and vice versa. To achieve both aims effectively would require two parallel, though potentially connected, evaluation processes. Alternatively, allocation, in particular accountability purposes, calls for transparency. This suggests that quantitative approaches may be most appropriate, and that there may be a high initial burden in setting up the framework.

Research evaluation tools typically fall into one of two groups, which serve different needs; multiple methods are required if researchers’ needs span both groups.

The two groups are:
• formative tools that are flexible and able to deal with cross-disciplinary and multi-disciplinary assessment
• summative tools that do not require judgement or interpretation, and are quantitative, scalable, transparent, comparable and suitable for high frequency, longitudinal use.

The units of aggregation used for collecting, analysing and reporting data will depend on the target audience(s), and the tool(s) used.

Developing a framework also requires careful selection of units of aggregation for the collection, analysis and reporting of data. The units selected for each evaluation phase are interrelated. The unit of reporting depends on the needs of the audience(s), as well as privacy requirements related to the unit being assessed. The unit of data collection will depend on feasibility, burden and the selection of tools, but must be at a lower (or the same) level of aggregation as reporting. The unit of analysis must be between (or the same as) those of collection and reporting.

There are some perennial challenges to research evaluation that need to be addressed.

The importance of these challenges varies depending on the aim(s) of the framework. Attribution and contribution are important when the evaluation is for accountability or allocation, but the fact that there may have been multiple inputs to downstream outcomes and impacts tends to be less important in advocacy, and may not be important in analysis depending on the purpose and context. Time lags between research being conducted and final outcomes are important when frameworks are using downstream measures of performance, such as outcomes and impacts. Being able to discriminate levels of performance at a fine-grained level is only really significant where the output of the evaluation process includes some kind of score or comparative measure between institutions (or groups, researchers and so on), typically for allocation purposes.

Research evaluation approaches need to suit their wider context.

Different approaches may be acceptable and credible in different environments, and it is important to consider this when developing a framework. The history, politics and wider social and economic context need to be taken into account. For example, bibliometrics as a methodology to evaluate research is credible and acceptable to the research community in some countries (e.g. Australia) but has met with
greater hostility in others (e.g. the UK). Credibility and acceptability can be improved by considering the risks of discrimination against specific groups (e.g. early career researchers, part-time workers and minority groups) through the implementation of the framework, and by carefully considering the unexpected consequences that may result. Moreover, the evaluation of emerging areas of research (for example, implementation science) presents special challenges and opportunities for new approaches, as traditional evaluation tools and metrics may be wholly unsuitable.

Finally, implementation needs ownership, the right incentives and support. Implementation is a key consideration, whether participation in the framework is compulsory or voluntary. Where compulsory, the challenge is to obtain support from the academic and wider community. Where participation is voluntary, incentives need to be in place to promote and sustain uptake. In both cases, participants need to be given the skills necessary for the process, through simplicity, training or a toolkit. In all cases, strong central ownership is needed for effective large-scale implementation.
Abbreviations

**CAHS**: Canadian Academy of Health Sciences. Developed a research evaluation framework using the payback framework, which is described in Appendix C.

**CIHR**: Canadian Institutes of Health Research. Major national research funder in Canada.

**ERA**: Excellence for Research in Australia. A research evaluation framework used nationally in Australia, which is described in Appendix C.

**ERiC**: Evaluating Research in Context. Research evaluation framework developed in the Netherlands, which is the precursor to the Productive Interactions approach.

**NIH**: National Institutes of Health. US federal health research funder.

**NIHR**: National Institute for Health Research. One national funder of health research in the UK.

**NSF**: National Science Foundation. US federal science funder.

**RAE**: Research Assessment Exercise. UK national research evaluation framework, which was discontinued in 2008. The predecessor of the REF.

**REF**: Research Excellence Framework. UK national research evaluation framework, which is described in Appendix C.

**RQF**: Research Quality Framework. Research evaluation framework developed in Australia but never implemented, which has provided significant input into the REF and ERA.

**STAR METRICS**: Science and Technology for America’s Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science. Research currently being developed in the US, which is described in Appendix C.

Other commonly used terms

**Characteristic**: used to define properties of research evaluation frameworks (and tools) that can differ between different evaluation units and against which they can be scored (e.g., ability to deal with multi-disciplinary research, comparability between institutions).

**Factor**: used to describe a group of characteristics that have been found through factor analysis to be (positively or negatively) correlated.
Chapter 1  Introduction

Research evaluation is used for multiple purposes: to provide accountability; for analysis and learning; to facilitate funding allocation; and for advocacy. There is a growing need to show clearly the impact of research and to provide evidence that publicly funded research presents a good return on investment for the research funders and, ultimately, the tax payer (Boehkolt, 2002; and Fahrenkrog et al., 2002). This requires a conceptual understanding of how research translates to academic, health, social and economic impacts, and how to measure those impacts through various forms of research evaluation.

The AAMC asked RAND Europe to conduct a detailed synthesis of existing and previously proposed frameworks and indicators used to evaluate research. The purpose of this review is to stimulate debate among US policymakers and the leadership of US medical schools and teaching hospitals on how best to measure the outcomes of research and ultimately account to US taxpayers and others for the returns on their investment in research. The review also aims to provide a guide to the key considerations that need to be taken into account when developing approaches to research evaluation, to be used as a resource by research funders, managers and policymakers.

The work builds on a previous RAND Europe review commissioned by the Canadian Academy of Health Research in 2008 (Brutscher, Wooding and Grant, 2008); see also CAHS (2009a). The aim of this current report is to update and develop the CAHS review. Specifically the objectives of the project are to:

- identify frameworks
- provide an overview and assessment of frameworks in use for research evaluation
- identify the research evaluation tools applied to those frameworks
- outline the considerations that need to be taken into account when developing a research evaluation approach, and the trade-offs that need to be made between different characteristics
- provide a guide to developing a research evaluation framework that can be used in a range of circumstances
- stimulate a wider debate on research evaluation in the US and internationally.

The structure and content of this report

To achieve the aims described above, we have reviewed 14 research evaluation frameworks, of which six were investigated in detail, and ten research evaluation tools. The six frameworks reviewed in detail were selected on the basis of relevance and interest to AAMC, and consist of a mix of national evaluation systems and health research specific approaches. The further eight frameworks were selected to cover frameworks which have not been reviewed in our previous work in this area, and to provide a wide geographic coverage, drawing on learning across a range of contexts internationally. The evaluation tools were selected to cover those used in the different frameworks described, and should cover the majority of approaches currently used in research evaluation. Detailed information on these frameworks is provided in Appendix C and the tools are described in Appendix D. These reviews are also summarised in the two tables in Appendix A, and a brief summary is provided below. Our subsequent analysis also draws on previous work look-
ing at other research evaluation frameworks. A list of these frameworks, with sources for additional information, is provided in Appendix E.

Having gathered these data on the frameworks and tools used and developed previously in research evaluation, we conducted a statistical analysis to understand how different characteristics of frameworks and tools interact. This consisted of identifying characteristics of research evaluation frameworks, and then scoring different frameworks against these characteristics. By looking at the correlations between these different characteristics we were able to establish some of the trade-offs and limitations that are inevitable in the development of approaches to research evaluation. A detailed description of the methodology used for this analysis is given in Appendix B. Drawing on these findings and the wider learning from the analysis of frameworks we explore the key considerations and challenges in developing approaches to research evaluation, illustrating how different aims, constraints and contexts can result in widely different approaches in different circumstances. This analysis and discussion is presented in Chapter 2.

Building on this, we have developed a guide to some of the key questions that should be addressed and challenges which should be considered in developing a research evaluation framework. This ‘how to’ guide is presented in Chapter 3, and gives an illustrative example of how it can be applied, using criteria which are likely to be of importance to the AAMC membership in developing a research evaluation framework.

Frameworks and tools reviewed

Frameworks
These are the frameworks selected for review in this study:

• Research Excellence Framework (REF): A framework developed to assess the performance of universities in the UK and to determine funding allocation, taking into account wider non-academic impacts of research.

• STAR METRICS: A framework under development in the US to assess and understand the performance of research and researchers, largely for accountability purposes, using data mining and other novel low burden methods. The approach so far only covers job creation (level 1) but is being developed to cover a wider range of research outcomes.

• Excellence in Research for Australia (ERA): A framework used in Australia to measure the performance of research in Australia, currently for accountability and advocacy purposes, and potentially for allocation in the future. The approach uses bibliometrics widely, as well as other quantitative indicators.

• Canadian Academy of Health Science (CAHS) Payback Framework: A framework developed in Canada using a logic model for health research translation, and drawing on a library of indicators. It aims to provide consistency and comparability between institutions in a research system with multiple regional funders.

• National Institute of Health Research (NIHR) Dashboard: A framework that consists of a dashboard to monitor the performance of research funded by the National Institute of Health Research in the UK, drawing on a logic model and a balanced scorecard approach. It accumulates data from a series of dashboards at lower levels of aggregations and is intended to be used for strategic decisionmaking and analysis.

• Productive Interactions: A framework developed across several countries in Europe and for multiple disciplines. It is a flexible approach which aims to help institutions learn and improve their performance against their own goals.

• Evaluation Agency for Research and Higher Education (AERES), France: Independent agency which evaluates French research and higher education institutions on the basis of internal and external assessments, with the aim of promoting learning and improvement.

• Congressionally Directed Medical Research Programme (CDMRP), US: Manages biomedical research for the US army medical research and material command through several processes, including a grant management system, product database and Concept Award Survey. The product database tracks products of research and their
stage of development against a number of criteria to map development and track where promising impacts and products are stalled or stopped.

- National Commission for University Evaluation and Accreditation (CONEAU), Argentina: A public body responsible for the evaluation of public and private universities and accreditation of their undergraduate and graduate degrees. Evaluation is based on self-assessment against institutional goals and supports performance improvement rather than comparative ranking.

- National Institute for Academic Degrees and University Evaluation (NIAD-UE), Japan: Conducts evaluations of teaching conditions and research activities at Japanese higher education institutions. Evaluation is against institution's research goals and includes a wider range of outcomes from research.

- Knowledge Management and Evaluation Directorate of the National Research Fund (NRF KM&E), South Africa: Monitors and evaluates research, covering the evaluation and quality rating of individual researchers and the review of programmes, disciplines and initiatives across the scientific disciplines (for external and internal stakeholders). The approach is focused on academic peer review.

- Performance Based Research Fund (PBRF), New Zealand: A tertiary education funding process, assessing the research performance of tertiary education organisations and funding them accordingly. The model has three elements: to reward and encourage the quality of researchers (60 per cent of the fund); to reflect research degree completions (25 per cent of the fund); and to reflect external research income (15 per cent of fund).

- Spanish State Programme for University Quality Assurance (PNECU), Spain: Issues annual calls for quality assessment in which either public or private universities can participate. There are two types of assessment projects: thematic projects, related to either one undergraduate degree or to a group of them within the same field of enquiry in one or various universities; and global projects, which cover one or various universities considered as a whole. Both projects are based on the assessment of the quality across teaching, research and university services management.

- Standard Evaluation Protocol (SEP), Netherlands: The national approach to the assessment of research institutions in the Netherlands, which assesses four aspects of research: quality (international recognition and innovative potential); productivity (scientific output); relevance (scientific and socio-economic impact); and vitality and feasibility (prospects for the future). Self-assessment is followed by a site visit and then an openly published assessment report. Institutions are ranked, but this does not directly determine funding allocation.

**Tools**

These are the tools selected for review in this study:

- bibliometrics: a range of techniques for assessing quantity, dissemination and content of publications and patents; uses quantitative analysis to measure patterns of publication and citation, typically focusing on journal papers

- surveys: provide a broad overview of the current status of a particular programme or body of research; widely used in research evaluation to provide comparable data across a range of researchers and/or grants which are easy to analyse

- logic models: graphic representation of the essential elements of a programme or process; aims to encourage systematic thinking and guide planning, monitoring and evaluation

- case studies: can be used in a variety of ways; flexible enough to capture a wide variety of impacts, including the unexpected, and can provide the full context around a piece of research, researcher or impact

- economic analysis: comparative analysis of costs (inputs) and consequences (outputs); aims to assess whether benefits outweigh opportunity costs and whether efficiency is achieved; generally, there are three types of economic analysis: cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA)

- peer review: review by peers, typically other academics in the same or a similar field, of outputs of research; rationale that subject experts are
• data visualisation: tool for data summarisation, presenting large amounts of data in a visual format for human comprehension and interpretation
• site visits: visit by evaluating committee to department and institution; generally consists of a series of meetings over one or more days with a range of stakeholders
• document review: review of existing documentation and reports on a topic.

uniquely qualified to assess the quality of the work of others

• data mining: allows access to and understanding of existing data sets; uses algorithms to find correlations and patterns and present them in a meaningful format, reducing complexity without losing information
• interviews: used to obtain supplemental information on areas of interest, generally to access personal perspectives on a topic, or more detailed contextual information
The purpose of this chapter is to understand the characteristics of different frameworks and tools, and how they correlate with one another. We observe that there are five clear relationships between the characteristics of frameworks:

- Quantitative approaches (which produce numerical outputs) tend to produce longitudinal data, do not require judgement or interpretation, and are relatively transparent, but they have a high initial burden (significant work may be required at the outset to develop and implement the approach).

- Formative approaches (focusing on learning and improvement rather than assessing the current status) tend to be comprehensive, evaluating across a range of areas, and flexible, but they do not produce comparisons between institutions.

- Approaches that have a high central burden (requiring significant work on the part of the body organising the evaluation process) tend not to be suitable for frequent use.

- Approaches that have been more fully implemented tend to have a high level of central ownership (by either the body organising the evaluation, or some other body providing oversight on the process).

- Frameworks that place a high burden on participants require those participants to have a high level of expertise (or should provide capacity building and training to achieve this).

We also find that tools fall into two broad groups, which are characterised by particular sets of characteristics:

- Formative, flexible and able to deal with cross-disciplinary and multi-disciplinary assessment

- Scalable, quantitative, transparent, comparable, free from judgement (after initial development) and suitable for high frequency, longitudinal use.

In this chapter, we explore these characteristics further, their interaction and the trade-offs they imply in research evaluation, and some of the wider issues around the process of research evaluation and the considerations required in the development of an approach to research evaluation.

Characteristics of research evaluation approaches

In assessing the six frameworks we investigated in detail, we identified a series of characteristics which were present in one or more of the frameworks. For example, some frameworks were formative, while others were more summative. Similarly, some were well designed to deal with challenges of assessing multi-disciplinary research, while others were focused primarily on health research. We identified 19 such characteristics across the six frameworks. Using a Likert scale, we scored the frameworks against the characteristics, to understand how they relate to one another, and then conducted a factor analysis. The details of the methodology used are described in Appendix B. These are the characteristics and measures we identified, with their definitions:

- Cross-disciplinary: the ability of the framework or tools to assess research that does not sit within conventional disciplines effectively and fairly

- Multi-disciplinary: applicability of the approach to a wide range of disciplines; some frameworks may be designed with a particular field in mind.
whether the approach is fully implemented and well established, or has not even been piloted

- **ownership**: of the evaluation framework by a central authority
- **free from judgement**: the extent to which the evaluation does not require the use of judgement and interpretation during the implementation phase
- **transparency**: the extent to which there are clear, transparent rules in relation to the approach and it is transparent to those external to the process
- **quantitative**: the extent to which the approach and the outcomes of the evaluation are highly quantitative, rather than qualitative.

It is useful to consider some of these characteristics in light of the rationales for research evaluation, as the choice of research evaluation frameworks to be used in different circumstances will be determined by the purpose of the evaluation, and the extent to which the characteristics of those frameworks meet those needs. As noted earlier, there are four key rationales for the evaluation of research. Any research evaluation is likely to be conducted for one or more of these purposes: advocacy, accountability, analysis and/or allocation. These rationales can be defined as follows:

- **advocacy**: to demonstrate the benefits of supporting research, enhance understanding of research and its processes among policymakers and the public; to make the case for policy and practice change
- **accountability**: to show that money has been used efficiently and effectively, and hold researchers to account
- **analysis**: to understand how and why research is effective and how it can be better supported, feeding into research strategy and decision-making by providing a stronger evidence base
- **allocation**: to determine where best to allocate funds in the future, making the best use possible of a limited funding pot.

It is important to note the distinction between advocacy and allocation, which may not be clear initially. Allocation is used where a decision needs to be made about which institutions, programmes or researchers to fund in the future, where that funding will
Challenges and trade-offs in research evaluation

are intended to be used for accountability purposes
are generally similar, with scalability and participant
burden important. Transparency is also a key con-
sideration, and this may be important for allocation.

If analysis is the purpose of the evaluation, a for-
formative approach is likely to be desirable. Depending
on the nature of the analysis required, it can also be
important that the approach is comprehensive.

Requirements for a framework which will be
used for advocacy are less clear, partly because the
characteristics needed are highly dependent on the
intended audience. This is a significant consider-
ation in the development of research evaluation
frameworks regardless of the purpose, and will be
discussed in more detail in the following sections.

Also important is the breadth and nature of
the research to be evaluated. A framework which
is intended to be used across the whole national
research system needs to be scalable, and be able
to deal with cross-disciplinary and multi-disciplin-
ary research. If the approach is intended to be used
within a single medical college, these considerations
will be less important.

We show in Table 2.1 how the six different
research evaluation frameworks we reviewed in
detail support and combine the different rationales.

It is interesting to note that for all approaches,
accountability is one of the rationales for the evalu-
ation, perhaps reflecting an increasing movement
towards increased accountability for the use of
public money in research. The other rationales are
more dispersed between the frameworks.

Understanding the rationale(s) of the research
evaluation framework is a crucial first step in under-
standing the way in which the evaluation should
be conducted, and has some clear implications in
light of the characteristics identified. For example,
frameworks which will be used for allocation usually
require some comparison between institutions as the
output of the evaluation process. Scalability and par-
ticipant burden are also likely to be important con-
siderations. Considerations for frameworks which
are intended to be used for accountability purposes
are generally similar, with scalability and participant
burden important. Transparency is also a key con-
sideration, and this may be important for allocation.

If analysis is the purpose of the evaluation, a for-
formative approach is likely to be desirable. Depending
on the nature of the analysis required, it can also be
important that the approach is comprehensive.

Requirements for a framework which will be
used for advocacy are less clear, partly because the
characteristics needed are highly dependent on the
intended audience. This is a significant consider-
ation in the development of research evaluation
frameworks regardless of the purpose, and will be
discussed in more detail in the following sections.

Also important is the breadth and nature of
the research to be evaluated. A framework which
is intended to be used across the whole national
research system needs to be scalable, and be able
to deal with cross-disciplinary and multi-disciplin-
ary research. If the approach is intended to be used
within a single medical college, these considerations
will be less important.

These relationships are summarised in Table 2.2.

Table 2.1
Purposes of the six frameworks investigated in detail

<table>
<thead>
<tr>
<th>Purpose</th>
<th>REF</th>
<th>ERA</th>
<th>Productive Interactions</th>
<th>NIHR</th>
<th>CAHS</th>
<th>STAR METRICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advocacy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Accountability</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Analysis</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Allocation</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

We show in Table 2.1 how the six different
research evaluation frameworks we reviewed in
detail support and combine the different rationales.

It is interesting to note that for all approaches,
accountability is one of the rationales for the evalu-
ation, perhaps reflecting an increasing movement
towards increased accountability for the use of
public money in research. The other rationales are
more dispersed between the frameworks.

Understanding the rationale(s) of the research
evaluation framework is a crucial first step in under-
standing the way in which the evaluation should
be conducted, and has some clear implications in
light of the characteristics identified. For example,
frameworks which will be used for allocation usually
require some comparison between institutions as the
output of the evaluation process. Scalability and par-
ticipant burden are also likely to be important con-
siderations. Considerations for frameworks which
are intended to be used for accountability purposes
are generally similar, with scalability and participant
burden important. Transparency is also a key con-
sideration, and this may be important for allocation.

If analysis is the purpose of the evaluation, a for-
formative approach is likely to be desirable. Depending
on the nature of the analysis required, it can also be
important that the approach is comprehensive.

Requirements for a framework which will be
used for advocacy are less clear, partly because the
characteristics needed are highly dependent on the
intended audience. This is a significant consider-
ation in the development of research evaluation
frameworks regardless of the purpose, and will be
discussed in more detail in the following sections.

Also important is the breadth and nature of
the research to be evaluated. A framework which
is intended to be used across the whole national
research system needs to be scalable, and be able
to deal with cross-disciplinary and multi-disciplin-
ary research. If the approach is intended to be used
within a single medical college, these considerations
will be less important.

These relationships are summarised in Table 2.2.

Table 2.2
The desirable characteristics of advocacy, analysis, accountability and allocation

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Desirable characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advocacy</td>
<td>Will depend on the audience</td>
</tr>
<tr>
<td>Analysis</td>
<td>Formative; comprehensive</td>
</tr>
<tr>
<td>Accountability</td>
<td>Comparability; scalability; participant burden; transparency</td>
</tr>
<tr>
<td>Allocation</td>
<td>Scalability; participant burden; transparency</td>
</tr>
</tbody>
</table>
The next section explores these characteristics and the relationships between them in more detail, investigating some of the trade-offs that take place in the selection of research evaluation frameworks and tools by analysing the correlation between these different characteristics in each case.

Analysis of framework characteristics
We know that there are trade-offs in the evaluation of research. Different frameworks are suitable for use in different circumstances and to meet different needs. There is no silver bullet or ‘ideal’ framework for the evaluation of research. In this section, we look at some of these trade-offs by analysing the characteristics of a range of different frameworks. To do so, we scored the 14 frameworks investigated, as listed in Chapter 1, plus an additional seven frameworks previously investigated in other work by RAND Europe, against each of the characteristics described in the previous section. Through a process of statistical analysis (described in Appendix B), we were able to identify correlations (both positive and negative) between these characteristics in these 21 frameworks.

The findings of this analysis are presented below. Although the relationships between the characteristics have been determined through this correlation-based analysis, the way in which the relationships are described, including the order in which factors are presented, has been determined drawing on prior experience and understanding of research evaluation processes. Therefore although the correlation between characteristics is determined quantitatively, the causation implied in the wording below draws on the knowledge and experience of a wide range of approaches to research evaluation, including the review of frameworks here and in previous studies (Brutscher, Wooding and Grant, 2008; and Grant et al., 2009). In all cases, the counter statement is also implied. For example, the first statement associates quantitative approaches with the production of longitudinal data. This also implies that qualitative approaches tend not to produce longitudinal data. The statements are presented in order of perceived importance, and are illustrated with examples from the six frameworks investigated in detail.

1. Quantitative approaches (those which produce numerical outputs) tend to produce longitudinal data, can be applied relative to fixed baselines reducing the need for judgement and interpretation, and are relatively transparent, but they have a high initial burden (significant work may be required at the outset to develop and implement the approach).

Example: Examples all relate to the inverse of this statement (which is also implied, as described above). A good one is the Productive Interactions framework, which is a qualitative approach with a low initial burden, but also requires a lot of judgement and interpretation and is not transparent. The approach does not tend to produce longitudinally comparable data. This is similar, though to a lesser extent, for CAHS Payback, and to some extent for the REF.

2. Formative approaches (which focus on learning and improvement rather than assessing the current status) tend to be comprehensive, evaluating across a range of areas, and flexible, but they do not produce comparisons between institutions.

Example: Productive Interactions is a formative approach, focused on learning and improvement, which is comprehensive and flexible but does not facilitate comparison between institutions. By contrast, ERA is a summative approach, which is rigid and inflexible, and gathers data on a specific finite set of indicators, but highly comparable between institutions. A similar comparison could be made between CAHS and STAR METRICS (Level 1).

3. Approaches which have a high central burden (requiring significant work on the part of the body organising the evaluation process) tend not to be suitable for frequent use.

Example: CAHS Payback has a high central burden in application and is applied relatively infrequently, in contrast to the NIHR Dashboard, which facilitates frequent data collection and has a low central burden. Productive Interactions is similarly infrequently applied and has a high central burden.

4. Approaches which have been more fully implemented tend to have a high level of cen-
Challenges and trade-offs in research evaluation

and multi-disciplinary assessment, or they are scalable, free from judgement, quantitative, transparent, comparable and suitable for high frequency, longitudinal use. The first group also tends to be more comprehensive, and the second group tends to have a higher initial burden. Therefore, to cover characteristics spanning the two groups a framework will likely require two or more complementary tools to be used. We can identify tools which fall into one or the other of these groups of characteristics:

- **formative, flexible and able to deal with cross-disciplinary and multi-disciplinary assessment:** examples: case studies, documentary review, site visits, peer review (to a slightly lesser extent)

- **scalable, quantitative, transparent, comparable, free from judgement and suitable for high frequency, longitudinal use:** examples: bibliometrics, economic analysis, data mining.

These relationships have implications for research evaluation for a number of different purposes. Frameworks which will be used for allocation are likely to require an output that facilitates comparison between institutions. This correlation analysis suggests that a framework that produces such comparative data is likely to be summative rather than formative, and is not likely to be comprehensive. This suggests that frameworks which are used for allocation are not likely to be well suited to use for analysis, and vice versa. Allocation purposes, and in particular accountability purposes, call for transparency. This suggests that quantitative approaches may be best suited to these purposes, but that this is likely to be at the expense of a high initial burden in setting up the framework. The first two factors in particular produce a space against which we can map the different purposes of research evaluation, as illustrated in Figure 2.1.

**Analysis of tool characteristics**

A similar analysis was conducted which allowed us to identify two groups of characteristics which are typically correlated for research evaluation tools. Two characteristics were excluded, since they are not relevant in the case of tools: the extent of implementation and the level of ownership. The methodology is described in more detail in Appendix B.

Our analysis suggests that tools are either formative, flexible and able to deal with cross-disciplinary

### Figure 2.1

Comparing characteristics and purposes of research evaluation frameworks

<table>
<thead>
<tr>
<th>Transparent, longitudinal, quantitative, free from judgement</th>
<th>Low up-front burden, qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formative, flexible, comprehensive</td>
<td>Analysis</td>
</tr>
<tr>
<td>Produces comparison</td>
<td>Advocacy</td>
</tr>
<tr>
<td>Allocation</td>
<td></td>
</tr>
</tbody>
</table>
developing an effective analysis, so tools from Group 2 are also likely to make a useful contribution. This underlines the importance of selecting an appropriate set of tools for evaluating research; multi-method approaches are most likely to be robust and serve the purpose of the evaluation most effectively. We investigated this further by looking at the combination of tools used in the six frameworks investigated in detail (Table 2.3).

Mapping the tools used in the six frameworks reveals that all frameworks make use of a range of tools. It is unsurprising to note that STAR METRICS uses only tools which fall into the second group of characteristics. The REF draws on tools from both groups, which perhaps reflects the fact that the approach is attempting to produce a comparison between institutions and be comprehensive, characteristics which are identified as being poorly aligned in the analysis of frameworks. This is similar, to a lesser extent, in the case of ERA, which draws on a fairly limited toolkit. The NIHR Dashboard relies primarily on tools with Group 2 characteristics, reflecting the fact that the approaches intended

These tools can be mapped onto axes representing the extent to which they have Group 1 and Group 2 characteristics, as illustrated in Figure 2.2.

Again this has implications for the way in which different tools can meet the needs of different research evaluation processes. Allocation and accountability purposes are likely to be achieved most effectively using tools having characteristics in the second group. However, where the framework is to be used across multiple disciplines, it is likely tools from the first group will also be required. Analysis purposes are likely to be well served by tools with characteristics from the first group. However, being free from judgement and interpretation can be important for developing an effective analysis, so tools from Group 2 are also likely to make a useful contribution. This underlines the importance of selecting an appropriate set of tools for evaluating research; multi-method approaches are most likely to be robust and serve the purpose of the evaluation most effectively. We investigated this further by looking at the combination of tools used in the six frameworks investigated in detail (Table 2.3).

Mapping the tools used in the six frameworks reveals that all frameworks make use of a range of tools. It is unsurprising to note that STAR METRICS uses only tools which fall into the second group of characteristics. The REF draws on tools from both groups, which perhaps reflects the fact that the approach is attempting to produce a comparison between institutions and be comprehensive, characteristics which are identified as being poorly aligned in the analysis of frameworks. This is similar, to a lesser extent, in the case of ERA, which draws on a fairly limited toolkit. The NIHR Dashboard relies primarily on tools with Group 2 characteristics, reflecting the fact that the approaches intended

---

1 Tools are plotted on two axes, representing their alignment with Group 1 and Group 2 characteristics. This was calculated by summing the scores for each tool against all the characteristics in Group 1, and dividing the number of characteristics in Group 1, to produce a number representing alignment with Group 1, and then conducting the same process from Group 2. The two scores are plotted against each other for each tool.
Challenges and trade-offs in research evaluation

The CAHS framework reinforces the importance of understanding the different levels at which frameworks are employed. Frameworks can use different units of data collection, and these may not necessarily correspond to the unit of reporting or the unit at which data are analysed. Some relationships between these different units which will exist within the same framework can be identified.

First, the unit of reporting must be at the same or a higher level of aggregation than the unit of analysis — logically, it is not possible to report data at a lower level of aggregation than that at which the analysis is conducted. Similarly, the unit of analysis must be at the same or a higher level of aggregation as the unit of collection for logical reasons.

When deciding what unit of aggregation to use for each of these three stages (data collection, reporting and analysis), the unit of reporting should be the first consideration. The unit will be determined largely by the intended use of the data. There may be multiple units of reporting corresponding to multiple audiences or multiple intended uses of the data. Confidentiality and data protection may also be a consideration: reporting in the public domain would most likely not include profiles at the individual to gather longitudinal data with frequent data collection. The Productive Interactions approach primarily uses tools from the first group of characteristics, reflecting the purpose of the framework, which is to be formative and flexible. Data mining has also been explored for use in relation to this framework, but this has not been well established and not integrated as a tool for use with this framework as yet.

CAHS Payback is particularly interesting as it draws primarily on tools that fall into Group 2, suggesting it is likely to be a scalable and comparable approach, rather than formative and flexible, whereas the framework is actually identified as comprehensive, flexible and formative rather than comparable. This may reflect the different levels at which CAHS can be employed. For example, the CAHS approach can be adapted and indicators selected to be specific contexts and to reflect the needs of different institutions, groups of institutions or bodies. However, once the tailored framework is developed, if applied in a consistent way it would produce comparable, transparent results independent of personal judgements for those using that specific approach. Thus the characteristics of the framework are in contrast to the characteristics of the tools it uses.

The CAHS framework reinforces the importance of understanding the different levels at which frameworks are employed. Frameworks can use different units of data collection, and these may not necessarily correspond to the unit of reporting or the unit at which data are analysed. Some relationships between these different units which will exist within the same framework can be identified.

First, the unit of reporting must be at the same or a higher level of aggregation than the unit of analysis — logically, it is not possible to report data at a lower level of aggregation than that at which the analysis is conducted. Similarly, the unit of analysis must be at the same or a higher level of aggregation as the unit of collection for logical reasons.

When deciding what unit of aggregation to use for each of these three stages (data collection, reporting and analysis), the unit of reporting should be the first consideration. The unit will be determined largely by the intended use of the data. There may be multiple units of reporting corresponding to multiple audiences or multiple intended uses of the data. Confidentiality and data protection may also be a consideration: reporting in the public domain would most likely not include profiles at the individual

Table 2.3
Cross mapping of tools used to six evaluation frameworks

<table>
<thead>
<tr>
<th>Tool</th>
<th>REF</th>
<th>ERA</th>
<th>Productive Interactions</th>
<th>NIHR</th>
<th>CAHS</th>
<th>STAR METRICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliometrics</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>
</tr>
<tr>
<td>Interviews</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document review</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data mining</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Data visualisation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Peer review</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economic analysis</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Surveys</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic models</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>
</tr>
</tbody>
</table>
theoratically the approach could be applied across the full range of levels of aggregation for data collection and reporting.

**Stage of measurement**

In research evaluation, measurements can be taken at various stages of the research process. Measures at four stages are typically identified reflecting steps through the logic model from research funding to wider benefit, as follows:

- input measures, capturing the resources consumed in the implementation of an intervention
- output measures, comprising the goods and services directly produced as a consequence of an intervention
- outcome measures, reflecting the initial impact of an intervention providing the reason for a programme
- impact measures, capturing the long-term changes an intervention brings about (Brutscher, Wooding and Grant, 2008; and CAHS, 2009a).

Less frequent, but also possible, is measurement of the ‘process’ step between input and output measures. This tends to be a measure used in frameworks which are intended for analysis purposes. It would be expected that the measures used will reflect the purposes of the evaluation process. Previous work (Brutscher, Wooding and Grant, 2008; and CAHS, 2009a) suggests that accountability and advocacy objectives tend to be associated with ‘upstream measures’ (outputs and outcomes), whereas ‘steering’ and ‘learning’ objectives tend to be associated with ‘downstream measures’ (outcomes and impacts, as well as process measures), as a result of the time lags between research being conducted and the more downstream measures materialising. However, this does not appear to be a clear finding for the six evaluation frameworks explored in detail in this study, as illustrated in Figure 2.4.

All the approaches are able to reflect output measures, and the majority also include some input measures. Outcome and particularly impact are more difficult to measure and this is reflected in these frameworks, with only two, REF and CAHS, including measures of all four types. However, there does
Challenges and trade-offs in research evaluation

13

It can be particularly difficult to place work that falls close to the boundary between the classifications, where the differential is small but the potential implications can be great for the researchers, programmes or institutions being evaluated. How can evaluation approaches deal with the incremental differences between research that falls at the boundaries between classifications?

- **Data quality**: All evaluations are necessarily limited by the quality and range of available data. Sources can differ in reliability, validity and accuracy and this inevitably affects the outcomes and reliability of the evaluation process.

- **Attribution (or contribution)**: Much research is incremental, and most research takes place within a community of researchers across institutions working on similar topics. Even where research is clearly differentiated from other work in the field, often other discoveries or progress in other fields are necessary to achieve impacts in the real world. How do we determine the extent to which a particular piece of research contributed to a particular impact? (Georghiou, 2002).

- **Time lags**: It can take a significant amount of time for research impacts to be realised, particularly in the case of basic research. Estimates in the field of medical research put the delay from research publication to uptake in healthcare practice at around 17 years (Morris, Wooding and Grant, 2011). How do we evaluate the impacts of research if we have to wait more than ten years for them to happen?

- **Examining the margin**: Often in comparative evaluation, we are looking to place research into classifications (e.g. excellent research vs good research, or transformative research vs incremental research). It can be particularly difficult to place work that falls close to the boundary between the classifications, where the differential is small but the potential implications can be great for the researchers, programmes or institutions being evaluated. How can evaluation approaches deal with the incremental differences between research that falls at the boundaries between classifications?

- **Data quality**: All evaluations are necessarily limited by the quality and range of available data. Sources can differ in reliability, validity and accuracy and this inevitably affects the outcomes and reliability of the evaluation process.

Attribution is an issue common to many of the frameworks. Where the planned use of the outputs of the evaluation are in the areas of analysis and advocacy, this may be less significant – it may be sufficient to be able to make the connection between research and its results. However, this will depend on the specific analysis and advocacy purposes. In some cases attribution can be important for analysis. For example, if you are analysing comparative investments, being able to measure outcomes and hence the return on your investment effectively is crucial. If you are only conducting an analysis in order to support learning and development within in a particular institution, however, this is likely to be less significant. More challenging are frameworks which aim to provide accountability and support allocation, as inevitably there will be ‘double counting’ of impacts where there were multiple or shared
Examining the margin is only really significant where the output of the evaluation process includes some kind of score or comparative measure between institutions (or groups, researchers and so on). Here, no formal approaches are stated in any of the frameworks discussed, and where this leads to allocation, this could prove challenging. The REF addresses this to some extent through its use of the criteria of ‘reach’ and ‘significance’ to assess impact case studies. For example, an advance in treatment which has been implemented in one hospital (perhaps where the research was conducted) might have great significance to a small number of people, so although significance is great, reach might be small. By contrast, an incremental development which has reached all the people that it could be relevant for may have a low level of significance but a large reach. By breaking down these two dimensions, it makes it easier to identify the truly outstanding case studies which have both great reach and great significance.

Time lags are only really an issue when frameworks are trying to evaluate the more downstream measures. Publications tend to occur relatively soon after research is conducted (compared to more downstream impacts, for example changes to healthcare practice) so when looking at purely academic outputs, this is a less challenging issue. When trying to understand wider downstream societal impact this does become a particularly challenging problem (Brutscher, Wooding and Grant, 2008; and CAHS, 2009a). REF addresses this by separating the assessment of impact and quality (academic excellence) and setting longer timescales for the assessment of impact. In the current round of the REF, due to be evaluated in 2014, case studies based on impacts from the period 2008–2014 will be eligible, but can be based on research dating back as far as 1993 (HEFCE et al., 2011a). This presents some further challenges, however, as it is likely that academics will have changed institutions since 1993, so attribution of the impacts of a particular research to a particular institution becomes challenging. For the REF, clear rules are in place, but this still remains a difficult issue. Productive Interactions also provides an interesting approach to this problem. Productive Interactions is used as a proxy for impact, assuming it is the route to impact, so it side steps the issue of time lag to impact. However, this requires the assumption that Productive Interactions and impact are directly correlated (Spaapen, et al. 2011).²

Data quality is a particular challenge in approaches that draw on existing data sets, such as data mining, but even more qualitative approaches such as case studies are reliant on appropriate triangulation and validation of data collected. The impact case studies used in the REF are also a good example in this case, where great care has been taken to ensure that appropriate evidence is required to demonstrate impacts claimed and their link to underlying research. This may be a particular challenge for Level 2 of STAR METRICS, which aims to use existing data sets. This makes the approach inherently reliant on the quality and relevance of data sets that are already available. Effective validation approaches will need to be developed to ensure the outcomes of the evaluation are credible and meaningful.

Credibility and acceptability

It is also important to take into account the credibility of frameworks, which can have a significant implication on their effectiveness. For example, the extent to which frameworks have a strong theoretical underpinning could have an impact on their credibility, particularly within the academic community. The NIHR Dashboard, for example, draws on a strong theoretical underpinning, which makes

² See http://www.eric-project.nl/nwohome.nsf/pages/NWOA_7RZC9Q.
it robust and credible (El Turabi et al., 2011). Developing the framework in collaboration with a range of stakeholders may also build credibility, as may building on approach, which has already been tested and showed to be effective elsewhere.

The acceptability of frameworks, particularly with research funders and other assessing bodies, but also with the academic community, also depends on the extent to which they produce perverse incentives. By introducing an assessment or evaluation process, there is always a risk that behaviours are changed in less desirable ways, which were unintended (Elton, 2000; Geuna, 2001; and Kostoff and Geisler, 2007). Perverse incentives can be manifested at several levels, from the institution to the research group or department, to the individual researcher. For example, if frameworks are not able to assess cross-disciplinary research effectively, then institutions may be reluctant to support the development of cross-disciplinary research efforts. At the group level, if frameworks are discriminatory against early career researchers, groups or departments may be more likely to employ experienced researchers, and reluctant to give early career researchers an opportunity. At the individual researcher level, if frameworks reward by volume of output, researchers may be tempted to subdivide research that could readily be combined into separate publications to boost their total number of publications artificially.

This risk is minimised in approaches such as Productive Interactions, where the assessment is made directly on the basis of the aims and mission of the individual groups under assessment.3 More flexible approaches like this are less susceptible to producing perverse incentives, while rigid and comparable approaches are more prone to producing these undesirable changes in behaviour. Similarly, the implications of the evaluation will affect the extent to which they produce perverse incentives. Frameworks used for funding allocation are more likely to lead to perverse incentives as there is more at stake in the outcomes of the evaluation. There has been significant discussion around this in the literature in relation to several frameworks (see for example Elton, 2000; Bradbent, 2010; Frolich et al., 2010; Kapeller, 2010; and van Vught and Westerheijden, 2012), and this was part of the discussion that led to the replacement of the Research Assessment Exercise (RAE) in the UK.

Frameworks also need to have credibility on a political level. A notable example is the case of Australia, where a change in government led to the Research Quality Framework (RQF) approach, which had been successfully piloted, being abandoned in favour of ERA. The key issue here was impact, which was extremely unpopular in particular with the research intensive institutions in the country, and heavy lobbying had led to the promise to abandon the RQF forming part of the election debate. It is interesting to note that in the UK, the REF drew heavily on the RQF when it was developed, but by contrast, while the impact element was preserved, the high level of use of bibliometrics was abandoned because of a lack of credibility of bibliometrics in the UK, and the high esteem in which peer review is held in the country (Ismail, Farrands and Wooding, 2009). Clearly, credibility of different approaches differs between countries and groups and depends on the requirements of the assessment approach, and the aims and outcomes of the assessment.

Another important element in developing a credible and acceptable framework is appropriate consideration of ways in which the framework might discriminate against different groups. Most typically these are early career researchers, but others are researchers who have suffered from illness, who have been on maternity or paternity leave, different racial or cultural groups, particular fields of research (which may be less mature, or may operate in ways that do not fit the ‘traditional model’, for example where publication behaviour is different) or particular types of institution (for example, centres that focus on cross-disciplinary research, or institutions that only have a short history of conducting research). Frameworks differ in the extent to which they address these issues. The REF has been very careful to address any concerns around discrimination that could result from the framework if applied without modification to all researchers (HEFCE et al., 2012). The number of publications required per

3 See http://www.siampi.eu (last accessed 12 August 2011).
researcher is reduced for researchers in a range of circumstances, with clear rules and processes laid out. To some extent, more flexible frameworks can deal with differing circumstances more readily; however, this will depend on sensitive and appropriate application of the evaluation approach and is at the discretion of individual departments and evaluators. For more rigid and/or quantitative approaches, there is a need to modify the framework. Where this is not done, the risk of discriminatory outcomes is higher than for more flexible approaches. In any case, clearly transparent approaches are critical.

Frameworks also need to be able to deal with the diversity of research that is conducted in health research, from hard lab-based science to the social sciences. If they are not carefully designed, frameworks may be better suited to different types of research, or may discriminate against some types of research. For example, bibliometric approaches are found to be suitable for application to the biological sciences, but are less appropriate for use in social sciences research because of differences in research outputs, journals and approaches to citation. A framework that was heavily reliant on bibliometrics and made direct comparison between these two areas without caveats or adjustment may well be discriminatory against the more social science oriented research areas.

Challenges of implementation

Even where frameworks have a high level of credibility across stakeholder groups, there can still be significant challenges in implementation. Some of these, particularly around expertise required of and burden on participants and centrally, and around issues in relation to ownership, are explored in the analysis above. However, there are many other factors to be taken into account. Participation is a crucial issue. In some cases participation is compulsory, which has the advantage of ensuring a complete data set, but it can reduce buy-in and provoke resentment and criticism. Where participation is not compulsory, uptake depends on a range of factors. Primarily, participants need to be able to see a clear benefit to themselves resulting from participation, and they need to have the ability to implement the framework. This again is dependent on burden and expertise, but can be supported by a clear explanation of the framework, which may include some kind of toolkit for implementation, or training from central owners of the framework. This has been noted as something that is lacking in the case of the CAHS Payback Framework, which has not been taken up widely to date, although it is early and uptake may increase over time. The complexity of implementation has also been criticised in the case of the Productive Interactions framework, and the study team has attempted to produce a toolkit to try and address this.

The future of research evaluation

Interest in and demand for the evaluation of research is increasing internationally. This is linked to a growing accountability agenda, driven by the demand for good governance and management growing in profile on national and international stages and fiscal austerity in a number of countries. There is a need to show that policymaking is evidence based, and in the current economic climate to demonstrate accountability for the investment of public funds in research. This is also complemented by a shift in emphasis from purely summative evaluations, which have traditionally characterised the assessment of research, to more formative evaluations, as well as more comprehensive evaluations that cover wider outputs from research outside the standard measures such as numbers and quality of publications.

In this environment of increasing numbers of evaluation processes and widening evaluation criteria, there is also a growing issue around standardisation. Many of the reasons for evaluation would be well supported by the ability to compare the outputs of different evaluation processes, from analysis, which would benefit from a larger data set to explore best practices, to accountability, which would benefit from greater comparability. In Canada, the Consortia Advancing Standards in Research Administration Information (CASRAI; casrai.org) is aiming to provide a centralised forum for researchers to standardise formats and semantics for sharing and aggregating research assessment and management data to ensure quality and comparabil-
Challenges and trade-offs in research evaluation

ity. This is particularly important in Canada, where there is a range of provincial funders, as reflected in the CAHS Payback Framework, which was developed in part to provide a library of standard indicators which could be used flexibly by the diversity of research funders across Canada to provide some comparability on a national scale, which had previously not been present.

The challenge with the development of standards, as has been found in the initial implementation of the CAHS Payback approach, is the tension between the use of standard indicators, and the flexibility to introduce ongoing new indicators and dimensions to assessments. The intention with the CAHS Payback approach is to provide a library of indicators to which new ones can be added as required. However, without the necessary central ownership to manage and develop this library, the uptake of this framework has been delayed (though it is growing). Reconciling flexibility and adaptability with standardisation requires regular central input and management.

Another key area that is growing in importance is data mining. As more research administration processes move online, making data accessible to data mining, and as data mining techniques develop, the research evaluation community is increasingly interested, as reflected by the development of STAR METRICS. This approach offers the potential to reduce significantly the burden of evaluation, but as yet it is largely untested. It is likely that there will be much iteration of such approaches, many ending in failure in the first instance, but nonetheless data mining has the potential to be a paradigm shift in the way research is evaluated.
Considering the characteristics, the trade-offs between them, and the challenges identified in our analysis of existing frameworks and research evaluation tools, we can start to put together a checklist of key issues that need to be considered in developing a research evaluation framework. This is presented below as a guide, outlining the key considerations that need to be taken into account. The order in which these considerations are presented is not intended to be a series of discrete, ordered steps; rather this is likely to be an iterative process, as there are significant interactions between the different decisions that need to be made. The one exception to this is the purpose of the evaluation, which should be determined at the outset, as it will feed into all the other considerations that need to be made.

To illustrate the approach presented, we also include a description of how this could be applied to develop a framework that may be suitable for use by a medical college in the US or Canada. We have developed a draft list of criteria that may be appropriate for a research evaluation framework to be used in this context. Some of these criteria would require further discussion and the development of any such framework would require discussion and consultation with the relevant medical college(s) to see how an evaluation approach could be developed to best meet their needs, mission and culture. Equally, there may be other criteria not included here which medical colleges may feel are important, and these would also need to be identified through a consultation process. The draft aims and criteria are set out in Box 1.

**Box 1:** Examples of likely aims of and criteria for a research evaluation framework for an American or Canadian medical school or teaching college

- The aim of the framework would be to provide an evaluation process that medical colleges can use to evaluate their research to promote learning and development, and quality improvement. The primary aim of the evaluation would be analysis, with accountability also important.
- The framework should be able to draw out academic and wider social, economic and policy impacts.
- The level of administrative burden, centrally and for participants, should be minimised.
- The framework should be transparent, with the rules and processes involved clear and unambiguous.
- The approach should avoid creating perverse incentives.
- The approach should be applicable to team-based research.
- The approach should be applicable to collaborative (including cross-disciplinary and cross-institution) research.
- The approach should support capacity building and the development of the next generation of researchers.
- It may be helpful if the framework could gather longitudinal data to support institutional and programme quality improvement.
- Medical colleges may prefer an approach that is not focused on producing rankings or comparisons between institutions.
Purpose

Question 1: What is the purpose of the framework?
The first step in the development of any framework is to have a clearly defined primary aim or purpose for the framework. This is the starting point for answering any of the questions in the following sections. This should define what the framework intends to achieve, considering all the stakeholders which will use it, and for what purpose. This may also include some ideas about the kinds of outputs that would be desirable for the framework. For example, allocation and accountability are best served by upstream measures of the inputs and outputs of research, whereas more downstream outcomes and impacts may be better suited to analysis and advocacy needs. A framework can have more than one aim but it should be clear about the primary purpose; having multiple aims may mean that a framework needs multiple sets of tools.

Example: The purpose of the framework is suggested above. It will be primarily for analysis, although it would be beneficial if accountability purposes are also served. This suggests that a range of downstream and upstream measures are likely to be required to meet both these needs. The audience will be internal researchers and research managers looking at ways in which they can improve and develop, and research funders, boards of trustees and the wider public in providing a demonstration of accountability.

Characteristics of the framework

In developing the framework, there are a number of trade-offs relating to the characteristics which need to be made, investigated below.

Question 2: What is the priority for your framework: to be independent of personal judgement and interpretation and transparent, gathering longitudinal, quantitative data, or to have a low initial burden and draw largely on qualitative data?

Our analysis suggests that transparent approaches which are free from judgement and can gather longitudinal data tend to be quantitative in nature and have a high initial burden. If you want your framework to be transparent you will need to take time at the outset to develop and hone the processes, and if you want your data to be independent of judgement and interpretation and comparable over time, this is more likely to be achieved by using quantiative data.

Example: Looking at the characteristics outlined in the box above, it is likely that a transparent framework, which gathers longitudinal, quantitative data, would be appropriate in this context. Therefore, there may be some initial burden in developing the approach.

Question 3: What is the priority for your framework: to be formative, flexible and comprehensive, or to produce data that allow comparison between institutions or programmes?

Our analysis suggests that formative, flexible, comprehensive frameworks do not usually produce data that facilitate comparison between institutions or programmes. If you want to compare institutions or programmes, but also have a formative element to your framework, it is likely that you will need to develop two parallel approaches to meet these conflicting needs.

Example: The production of comparisons between institutions is considered undesirable here, so the priority for the framework would be to be formative, flexible and comprehensive.

Question 4: Do you want a framework that can be applied frequently?

Our analysis suggests that frameworks that can be applied frequently have a low level of central burden (for the implementing body), most likely because of issues of feasibility. If you have a complex framework, which requires a lot of central input, it is unlikely that it can or will be applied frequently.

Example: It is likely that a framework that can be applied reasonably frequently would best support quality improvement and performance management, so if that was considered desirable, the framework would need to have a fairly low central burden.

Question 5: Do you have the ownership you need to reach the level of implementation you require?

Our analysis suggests that a high level of central ownership is associated with a high level of imple-
Building a new research evaluation framework

21

in this area. Attribution and contribution is less critical from an analysis perspective, but if accountability is an aim of the framework it will be important to use clearly defined processes for the assessment.

Consider the context

Whether your developing framework is credible and acceptable will depend on the context in which it is to be applied.

Question 7: Who are your key stakeholders?
It is important to establish to whom the framework needs to be credible and acceptable. Possible groups to consider include the academic community, funding bodies, charities, boards of governors and trustees, special interest groups, relevant government departments and the public. Each will have different views on what is credible and acceptable.

Example: Key stakeholders include the medical college(s), their board(s) of trustees, the relevant research funders and the general public. Funders in the public sector will most likely be predominantly the National Institutes of Health (NIH) in the US along with specialist subject area funders (e.g. the National Institute of Mental Health), and Canadian Institutes of Health Research (CIHR) in Canada, along with a wide variety of regional research funders. There may also be charitable and private sector funders.

Question 8: How do you make the framework credible to the key stakeholders?
Some key issues to consider are the potential for generating perverse incentives, the theoretical underpinning of the framework, and the extent to which the framework may discriminate or be perceived to discriminate against different groups. It is also important to take into account the cultural and historical context. For example, what has been tried before, and what was the critical reception? What tools are considered acceptable? Different approaches will be acceptable in different countries and communities, and for different purposes. Communication is crucial.

Example: Medical colleges as participants will likely wish to avoid approaches that have a high burden, and will also need to take into account any
possible consequences of the evaluation for them, particularly those related to funding and reputation. It is important that the approach developed does not discriminate against early career researchers or other groups, and this suggests that quantitative tools should be used with care, especially at the specific unit level. The NIH awards most of its funding through peer review, so this is clearly an acceptable approach to the assessment of research in the region. However, the STAR METRICS project indicates that there may be an appetite in the US for more novel approaches, and for metrics based on quantitative approaches, so there is some scope to try new and novel methods. Also, CAHS Payback and a similar payback approach at CIHR demonstrate that there is scope in Canada for indicator-based alternative approaches.

**Select the tools**

Tools tend to fall into one of two groups:

- tools that are formative can be flexible and are able to deal with cross-disciplinary and multidisciplinary assessment (examples: case studies, documentary review, site visits, peer review)
- tools that are scalable, quantitative, transparent, comparable, free from judgement and suitable for high frequency, longitudinal use (examples: bibliometrics, economic analysis, data mining).

To cover characteristics groups, you need to use tools from each group (see figures 2.1 and 2.2). It is also important to take into account the expertise of those employing the tools, centrally or among participants.

**Example:** According to the criteria laid out this framework would need to be formative and flexible, but also scalable, transparent and suitable for longitudinal use. This would require the use of tools from both categories.

**Level of aggregation**

**Question 9: What is your unit of reporting?**

The unit of reporting is the way in which the final outcomes of the evaluation are presented (e.g. if a report is to be published will this be broken down at institution, programme level or individual level?) This will depend on the way in which the data will be used. There may be multiple audiences for the results of the evaluation so it is important to consider the full range of possible applications. It is also important to consider privacy concerns of individual researchers.

**Example:** Reporting is likely to be on the level of the medical college, or perhaps programmes or research groupings within the school and its teaching hospital affiliate(s). It would be undesirable to report at a lower level of aggregation, such as at the individual researcher level. It may also be valuable for accountability purposes to provide reporting outputs at a higher level of aggregation, for example spanning a number of medical colleges.

**Question 10: What is your unit of collection?**

You will need to determine the level at which data will be gathered. By this, we mean will you request data, for example a list of publications for each individual, for each research group, by programme or across a whole institution? This will depend on the nature of the framework and the tools selected. The unit of collection must be at or below the unit of aggregation used for reporting.

**Example:** This will depend on the tools ultimately selected, but to reduce discrimination, and take better account of team-based working, it might be desirable to collect data at the project or small research group level.

**Question 11: What is your unit of analysis?**

This determines the level at which data are aggregated and analysed, and depends on the types of data, the purpose and characteristics of the framework, and the way in which the data will be used. For example, if you are conducting a bibliometric analysis, will you ask for an analysis of publications at the individual level, or will you aggregate this by research group and conduct the analysis at this level (or any other level)? The unit of analysis will necessarily fall at or between the units of aggregation used for reporting and collection.

**Example:** This will depend on the tools ultimately selected, but to reduce discrimination, and take better account of team-based working, it might
be desirable to collect data at the project or small research group level.

**Implementation**

**Question 12: Do you have ownership?**
As described above, central ownership is strongly associated with high levels of implementation. If possible, this should be ensured.

**Example:** It would be valuable if there were a central point to provide stewardship of the framework and support to medical colleges in implementation. This may be a role that AAMC or another relevant body could fill.

**Question 13: Is it compulsory?**
If the framework is compulsory, it is important to consider how you can ensure acceptability and buy-in. If participation is voluntary, it is important to consider how you can make it easy and desirable to participate and sustain engagement. Consider the incentives for participation, the level of participant burden, and whether providing information, through training or a toolkit, could be beneficial.

**Example:** There is no mandate for a compulsory framework, so it will be voluntary, and therefore could helpfully be supported by training and/or the provision of a toolkit to enable uptake. Incentives for participation need also to be considered. If it is possible that the information used could contribute to grant applications that would certainly support uptake.

**Summary: example case**
Based on the criteria suggested above, we now have developed an overview of how this approach could be applied to develop a framework that may be suitable for use by a medical college in the US or Canada. Drawing together these considerations, we suggest to institutions that an approach that draws on and builds on the type of indicator library developed in the CAHS Payback Framework combined with the NIHR Dashboard approach could be suitable. Drawing on the CAHS Payback approach would facilitate uptake in Canada and is likely to be acceptable in the region and to many funders, and familiar to some institutions. Increased central ownership and a repository for indicators and data would help facilitate uptake. The approach would require some initial effort to set up, but once developed institutions could draw on some of the formative methodologies used in the Productive Interactions approach, working internally to develop their own appropriate subset of indicators, building on a logic model such as those used in the CAHS Payback approach and the NIHR Dashboard, and then reflecting internally on their progress against these indicators and the underlying processes that could be improved and developed to promote better performance. Care needs to be taken to ensure this initial development process does not place too heavy a burden on researchers and participants.

The indicator set would need to be considered carefully to ensure there is no discrimination against early career researchers and other disadvantaged groups. This could be partly addressed by focusing data collection at the project or small group level, rather than the individual level, which would also help promote team-based research and collaboration. By building on a set of quantitative underlying indicators, which provide transparency, with an internal learning process, similar to a site visit, the framework could be clear and transparent, and formative and flexible, meeting both accountability and analysis needs. Since indicators would be selected by institutions as a subset of a wider list of possible indicators, the ability to compare longitudinally is preserved, without producing an overall comparison or ranking of performance between institutions. Consideration of other indicators used already in research administration may help reduce burden, and discussion with standards bodies such as CASRAI in developing the indicator set could support this.

The risk is that institutions may not have the expertise to utilise the framework and develop an appropriate set of indicators, but if the collection of data corresponding to the indicators can be made low burden, and the process by which the analysis of the outcomes and processes takes place simplified and explained through a toolkit, this should facilitate uptake. Central ownership would be helpful in promoting implementation, and in Canada the correspondence of the framework to the CIHR and
CAHS Payback approach may help provide incentives for uptake.

Incentives for participation in the US require further consideration. Careful consideration also needs to be given to how indicators can effectively address time lag between research and impact. It should also be noted that there have been acknowledged challenges in implementation of the CAHS Payback Framework, so implementation in particular will need careful consideration. Providing support in implementing the framework would overcome some of these challenges. Providing clear, straightforward guidance materials would also support this. Strong central ownership, with the support of a group or body which could act as a central repository for the indicator library, may help overcome some of these challenges. Monitoring pilot or demonstration projects would be a useful way to investigate and find solutions to some of these challenges.

**Decision tree**

We can combine some of the considerations outlined to form a decision tree, which demonstrates how some of the key decision steps inform the type of framework you can use (Figure 3.1). This does not comprehensively cover all the issues addressed in the sections above, so should be used in conjunction with a wider analysis of the needs of your evaluation and the context in which it takes place. It does, however, provide a useful starting point for thinking about some of the key fundamental decisions and their implications for the development of an appropriate framework.
Figure 3.1
A decision tree for developing a research evaluation framework

What is the purpose of the evaluation?

Analysis
- Up & downstream measures appropriate
- Formative so not likely to be comparable
- Need to consider time lags

Accountability
- Upstream measures appropriate
- Free from judgement and transparent, so quantitative and high initial burden
- Need to consider attribution

Advocacy
- Downstream measures appropriate
- No need to be frequent so can have high central burden
- May need to consider time lags

Allocation
- Upstream measures appropriate
- Comparison needed, cannot be formative, flexible, comprehensive
- May need to consider attribution

What type of research are you evaluating?

Level of aggregation: What unit of data reporting, collection and analysis will you use? Reporting ≥ analysis ≥ collection

Context: Who are your stakeholders? What will be credible and acceptable to them? What has been done before?

Implementation: Do you have strong central ownership? What burden does the framework place on participants, and how are they supported and incentivised to participate?

How many institutions are you evaluating?

Multi-disciplinary or cross-disciplinary
- Need tools from both groups

Multi-disciplinary or cross-disciplinary
- Can use any tools

Single-disciplinary
- Need tools from Group 1

Single-disciplinary
- Need tools from Group 2

One or few institutions
- Many institutions


HEFCE et al. (2011b) *Consultation on Draft Panel Criteria and Working Methods*, REF 03.2011, Higher Education Funding Council for England, Scottish Funding Council, Higher Education Funding Council for Wales,

Measuring research: a guide to research evaluation frameworks and tools


McAllister, P. R. and Wagner, D. A. (1981) ‘Relationship Between R&D Expenditures...


Starkey, S. et al. (forthcoming) Initial Deployment of the RAiSS Tool for the National Institute for Health Research, RAND Report PM-3702-NIHR.


UCLA Anderson School of Management (n.d.) ‘Data Mining: What is Data Mining?’ As


## Appendix A: Summary tables for frameworks and tools reviewed

### Table A1
Summary table of the frameworks reviewed in this study

<p>| Framework                           | Origin and rationale                                                                 | Scope                                                                 | Measurement                                                                                           | Application to date                                                                                     | Analysis                                                                                                    | Wider applicability                                                                 |
|------------------------------------|---------------------------------------------------------------------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| Research Excellence Framework (REF) UK | Evolved from its predecessor, the RAE, and the RQF. Intended to be low burden, but pressure from researchers led to changes. Includes wider societal impact. | Assessment at subject level on three elements: quality of research outputs, impact of research (not academic) and vitality of environment. | Assessment by subject peer review panel of list of outputs, impact statement and case studies, and statement on research environment. | Piloted 2009. First round of assessment 2014; results will determine funding allocation. | Burden not reduced, but adds wider impact to evaluation. Originally metrics based, but this was dropped as too unpopular. | Suitable for similar cross institutional assessment of performance. High burden on institutions, arguably expensive, so best for significant funding allocation uses. |
| STAR METRICS, US                   | Key aim to minimise burden on academics. Helps to meet US federal accountability requirements. | Two levels: Level 1, number of jobs supported; Level 2, range of research funded researcher interactions and wider impacts. | Data mining approach, automated. At present, only gathers jobs data. Methodologies for Level 2 still being developed. | Level 1 rolled out to 80 universities. Level 2 still under development. | Feedback generally positive, but feasibility of Level 2 not proven. | Potentially very wide depending on success of Level 2. There has been international interest, eg from Japan, EC. |
| Excellence in Research for Australia (ERA), Australia | Perceived need to include assessment of quality in block funding allocation (previously volume only). Advocacy purpose to demonstrate quality of Australian research. | Assesses quality, volume, application of research (impact), and measures of esteem for all Australian universities at disciplinary level. | Indicator approach; uses those appropriate at disciplinary level. Dashboard provided for review by expert panel. | First round in 2010, broadly successful. Next round 2012, with minor changes. Intended for funding allocation, but not used for this as yet. | Broadly positive reception. Meets aims, and burden not too great. Limitation is the availability of appropriate indicators. | Should be widely applicable; criticism limited in Australian context. Implementation appears to have been fairly straightforward. |</p>
<table>
<thead>
<tr>
<th>Framework</th>
<th>Origin and rationale</th>
<th>Scope</th>
<th>Measurement</th>
<th>Application to date</th>
<th>Analysis</th>
<th>Wider applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Academy of Health Sciences (CAHS) Payback Framework, Canada</td>
<td>Draws on well-established ‘payback’ framework. Aims to improve comparability across a disparate health research system. Covers wide range of impacts.</td>
<td>Five categories: advancing knowledge; capacity building; informing policies and product development; health and health sector benefits; broader economic benefits.</td>
<td>Specific indicators for each category. Logic model has four research ‘pillars’: biomedical; clinical; health services; social cultural, environmental and population health.</td>
<td>Used by public funders; predominantly CIHR (federal funder), but there has also been some uptake by regional organisations (e.g. Alberta Innovates).</td>
<td>Strengths: generalisable within health sector; can handle unexpected outcomes. But understanding needed at funder level – may limit uptake. Early stages hard to assess.</td>
<td>Breadth, depth and flexibility mean framework should be widely applicable. However, it only provides a guide and needs significant work to tailor to specific circumstances.</td>
</tr>
<tr>
<td>National Institute of Health Research Dashboard, UK</td>
<td>Aim is to develop a small but balanced set of indicators to support strategic decisionmaking, with regular monitoring of performance.</td>
<td>Data collected quarterly at programme level on inputs, processes, outputs and outcomes for three elements: financial, internal process and user satisfaction.</td>
<td>Programme specific data can be pooled to provide a system level dashboard; 15 indicators selected, matching core aims, collected quarterly.</td>
<td>Launched July 2011 NIHR-wide, with data to be provided by the four coordinating centres, analysed and aggregated centrally.</td>
<td>Designed to fit strategic objectives, so in that sense likely to be effective. However, only just launched, so detailed analysis premature.</td>
<td>Should be applicable to other national health research funders. Performance indicators selected can be tailored to assessment needs.</td>
</tr>
<tr>
<td>Productive Interactions, Europe</td>
<td>Measures productive interactions, defined as interactions with stakeholders that lead to change. Eliminates time lag, easier to measure than impacts. Assessment against internal goals intended for learning.</td>
<td>Intended to work in a wide range of contexts, best applied at research group or department level where goals are consistent.</td>
<td>Three types of interaction: direct personal contacts, indirect (e.g. via a publication) and financial. Engages users; findings assessed against internal goals.</td>
<td>Piloted across diverse disciplines and contexts in four European countries and at EC level. No plans to roll out more widely at present.</td>
<td>Tailored, so should help improve performance. Requires significant work from participants to generate their own set of goals and indicators.</td>
<td>Indicators developed to meet goals, so widely applicable, but does not produce comparison between institutions, so not appropriate for allocation, and could be challenging to use for accountability.</td>
</tr>
</tbody>
</table>
Table A2
Summary table of the characteristics of the tools reviewed in this study

<table>
<thead>
<tr>
<th>Tool</th>
<th>What is it?</th>
<th>When should it be used?</th>
<th>How is it used?</th>
<th>Group 1 or 2?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliometrics</td>
<td>Range of techniques for assessing quantity, dissemination, and content of publications and patents. Uses quantitative analysis to measure patterns of publication and citation, typically focusing on journal papers.</td>
<td>Advantages: low cost and burden, can map collaboration and inter-disciplinarity. Challenges: journal coverage; identifying author affiliations; choosing timeframe for analysis. Caveats: citation behaviour differs between fields and individuals and is not necessarily a good measure of quality. Do not cover wider impacts.</td>
<td>Measures activity (volume of outputs); knowledge transfer (citations used as a proxy for communication); linkage (links between individuals and research fields can be mapped); citation analysis (proxy for quality of research).</td>
<td>2</td>
</tr>
<tr>
<td>Case studies</td>
<td>Flexible enough to capture a wide variety of impacts, including the unexpected, and can provide full context around a piece of research, researcher or impact.</td>
<td>Primary limitation is the generalisability of findings – best used where examples are needed rather than full coverage. Often used to provide the contextual information alongside another approach which provides generalisable information and to understand process variables in depth.</td>
<td>Can be used in a variety of ways depending on aims. Key considerations are the unit of analysis and the sample selection for multiple case studies. Structure and approach should reflect the purpose.</td>
<td>1</td>
</tr>
<tr>
<td>Data mining</td>
<td>Allows access to and understanding of existing data sets. Uses algorithms to find correlations or patterns and present them in a meaningful format, reducing complexity without losing information.</td>
<td>Developing effective data mining processes might be complex and time consuming, but potential to reduce burden of data collection on researchers by using information already being collected. Key limitation is reliance on availability and quality of existing data.</td>
<td>Key steps: prepare the data (select and process – cleaning, transformation where necessary); develop, conduct and test the data mining processes; interpret the information (including clear presentation of data).</td>
<td>2</td>
</tr>
<tr>
<td>Data visualisation</td>
<td>Tool for data summarisation, presenting large amounts of data in a visual format for human comprehension and interpretation.</td>
<td>Can allow visual interpretation of large data sets. Particularly valuable when data are highly heterogeneous and noisy; allows data to be explored intuitively without mathematical processing.</td>
<td>Often used in combination with data mining; data mining identifies patterns and insights from data; visualisation used to present the data.</td>
<td>–</td>
</tr>
<tr>
<td>Document review</td>
<td>Review of existing documentation and reports on a topic.</td>
<td>Gives a broad overview of an issue; identifies ‘what is known’. Should be tightly focused to avoid time-consuming trawling for evidence.</td>
<td>Often a first step before other tools used. Can add to existing knowledge through synthesis despite collecting no new primary data.</td>
<td>1</td>
</tr>
<tr>
<td>Economic analysis</td>
<td>Comparative analysis of costs and consequences. Purpose is to assess whether benefits outweigh opportunity costs and whether efficiency is achieved. Three types: cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA).</td>
<td>A systematic way of producing comparisons. CEA and CUA should not be used if data on alternatives are not comparable, or it is necessary to understand value of one unit of research independent of comparison or externalities. CBA provides wider understanding of outcomes. Should not be used if assigning monetary value is inappropriate or impossible.</td>
<td>CBA: expresses impacts in monetary units, often most useful but also most difficult. CEA: compares outcomes between alternatives, and estimates of expected costs and outcomes are expressed in a single dimension measure. CUA: compares costs and benefits of alternatives to help determine worth relative to an external standard.</td>
<td>2</td>
</tr>
<tr>
<td>Tool</td>
<td>What is it?</td>
<td>When should it be used?</td>
<td>How is it used?</td>
<td>Group 1 or 2?</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Interviews</td>
<td>Used to obtain extra information on areas of interest, generally to access personal perspectives on a topic, or more detailed contextual information.</td>
<td>Advantages: detailed contextualised information; flexible; can cover a wide range of factors. Limitations: interviewers need to be skilled or can result in bias; time consuming; may not be generalisable.</td>
<td>Can be unstructured and open-ended (but require interviewer skill) or structured (easier to conduct and analyse but lose flexibility).</td>
<td>1</td>
</tr>
<tr>
<td>Logic models</td>
<td>Graphic representation of the essential elements of a programme or process. Aims to encourage systematic thinking and guide planning, monitoring and evaluation.</td>
<td>Limitations: do not capture the counterfactual; may change over time; often fundamentally linear. Advantages: provide a systematic structure to aid thinking; visual presentation; can represent a shared understanding.</td>
<td>Require inputs from other tools to populate logic model. Best established collaboratively, then tested and refined. May require several visual presentations for different audiences.</td>
<td>–</td>
</tr>
<tr>
<td>Peer review</td>
<td>Review by peers, typically other academics in the same or a similar field, of outputs of research. Rationale is that subject experts are uniquely qualified to assess the quality of the work of others.</td>
<td>Key advantage: credibility with the academic community. Criticisms: cost and burden; can be conservative; not transparent; can be slow; may disadvantage early career researchers. May be less valid where assessing wider outputs of research – may need to bring research users into the assessment process.</td>
<td>Materials reviewed by peers as a group or individually. Group review allows discussion, limits clarification by time, and gives opportunity to reach consensus, but is costly and requires effective facilitation. Individual review costs less but can be slow. Output may be a score, ranking (can be combined from several reviewers) or a qualitative review, with feedback.</td>
<td>1</td>
</tr>
<tr>
<td>Site visits</td>
<td>Visit by evaluating committee to department and institution. Generally consist of a series of meetings over one or more days with a range of relevant stakeholders.</td>
<td>Advantages: give access to a range of people, and opportunity for two-way feedback. Transparent to participants. Disadvantages: time consuming, and so not highly scalable; not suitable for frequent use and not transparent to external parties.</td>
<td>Data validation rather than gathering. No fixed process; generally a series of meetings with stakeholders, then discussion within evaluation panel. Typically follows initial reporting, (e.g. self-evaluation report), which informs the visit, and produces a final report, including assessment of current situation and advice for improvement.</td>
<td>1</td>
</tr>
<tr>
<td>Surveys</td>
<td>Provide a broad overview of the current status of a particular programme or body of research and are widely used in research evaluation to provide comparable data across a range of researchers and/or grants which are easy to analyse.</td>
<td>Tend to provide broad rather than deep information, but comparable between respondents and can cover a large sample. A limitation is lack of flexibility to adapt to individual circumstances, but this improves generalisability and reduces bias. Costs vary depending on approach and required response rate, but typically cost effective.</td>
<td>Can be employed in a range of ways, from online surveys to postal or telephone surveys. Methodology depends on purpose, but consists of the following steps: develop the survey (approach and question set); identify the sample; pilot the survey; conduct the survey; analyse findings.</td>
<td>2</td>
</tr>
</tbody>
</table>
This appendix provides a more detailed description of the methodology used in this study, covering the analysis of trade-offs and challenges in research evaluation. The approaches used in the review of existing frameworks and tools are described at the beginning of Appendix C and Appendix D respectively.

**Analysis of characteristics of evaluation frameworks**

The aim of this process was to try and understand some of the trade-offs that occur in the evaluation of research. Anecdotally, it is already known that such trade-offs exist, and that there is no ‘ideal’ framework for the evaluation of research. The aim of this analysis is to try and establish what patterns exist across frameworks and tools that have been developed and implemented (to differing extents) in the real world. Looking at these characteristics enables us to understand what trade-offs seem to exist on the basis of what has already been done. There are some important caveats, however. First, the data set is limited by the number of frameworks we have analysed, and is relatively small for this type of statistical analysis. Therefore, findings should be taken in a broad qualitative sense – we are looking at the nature of the relationships between framework characteristics, not trying to quantify them specifically. Second, this analysis only considers what has been done in practice – it may be theoretically possible to use frameworks in different ways or to generate frameworks that have completely different mixes of characteristics. We have only analysed what has been done in practice. Furthermore, this does not cover every research evaluation framework comprehensively, but it does cover a wide range of examples internationally.

**Identification of factors**

A strengths, weaknesses, opportunities and threats (SWOT) analysis was conducted by the project team on the six frameworks investigated in detail: Productive Interactions, REF, NIHR Dashboard, ERA, CAHS Payback and STAR METRICS. This was used to identify characteristics that affect the suitability of frameworks for use in different circumstances and for different purposes. The individual SWOT analyses are included in the framework reviews in Appendix C. A full list of the characteristics and definitions of these research evaluation approaches can be found in on pages 5–12.

**Mapping frameworks against dimensions**

A number of the characteristics identified could be expressed as dimensions against which the frameworks could be scored by RAND Europe experts in the evaluation of research. This was translated onto a Likert scale from 1 to 7 against which the six frameworks were scored by the project team during a workshop. The STAR METRICS approach was split into Level 1 and Level 2 for this process, according to the differences in scope and development of the two levels. The results of the mapping of frameworks against these axes are show in Table A3.
Since the number of characteristics identified is significantly greater than the number of frameworks considered in Table A3, it was necessary to include a wider range of frameworks in the analysis so that the subsequent statistical analysis is feasible. These frameworks were incorporated: French Evaluation Agency for Research and Higher Education framework (AERES), US Congressionally Directed Medical Research Programs framework (CDMRP), ‘Economic Impacts in Research and Innovation Framework’ of the UK Department for Innovation, Universities and Skills (DIUS), Evaluating Research in Context (ERiC), evaluation system for Framework Programme 7 of the European Union (EU), Leiden University Medical Centre framework (LUMC), Measure of Research Impact and Achievement (MORIA), Japan’s National Institution for Academic Degrees and University Evaluation framework (NIAD-UE), framework of the Knowledge Management and Evaluation Directorate of the South African National Research Foundation (NRF KM&E), RAND Assessment Impact Scoring System (RAISS), Australian Research Quality Framework (RQF), Dutch Standard Evaluation Protocol (SEP) and Swedish governmental agency for innovation systems (Vinnova). A number of them (AERES, CDMRP, NIAD-

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>SIAMPI</th>
<th>REF</th>
<th>NIHR</th>
<th>ERA</th>
<th>CAHS</th>
<th>METRICS Level 1</th>
<th>METRICS Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-disciplinary</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Longitudinal</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Comparability</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Multi-disciplinary</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Formative</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Flexible</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Participant burden</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Central burden</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Participant expertise required</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Central expertise required</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>Extent of implementation</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Scalability</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>Ownership</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Frequency</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Initial burden</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Free from judgement</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Quantitative</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Transparency</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>-</td>
</tr>
</tbody>
</table>
A number of additional characteristics and concerns in research evaluation frameworks were identified that were less suitable for scoring against in this way:

- whether the framework is sensitive to institutional goals of the evaluation
- whether the framework creates perverse incentives
- whether there is a toolkit for implementation
- how frameworks address the possible issues of discrimination (particularly in relation to early career researchers)
- what the best way to support implementation is (considering challenges of burden and expertise)
- what assumptions the approach makes
- how strong the theoretical basis is
- what the units of assessment, aggregation, analysis and reporting are
- what the levels of participation are, and how they differ between compulsory and voluntary evaluation
- how the context informs the development of a framework, considering issues such as the acceptability of different approaches, the surrounding politics and the ownership of the framework
- how the framework deals with the usual evaluation challenges, e.g. time lags, attribution
- how many methods the approach uses; what the value of multiple methods is
- how many indicators the framework uses, and how broad the assessment is; whether the indicators are balanced, selective or non-selective.

These were noted and considered in the wider discussion of the issues surrounding the development of approaches to research evaluation.

Correlation analysis

Plotting the characteristics using a scatter plot matrix (Figure A1) reveals that the data are not linearly related. Therefore, it was necessary to convert to scores to rankings before correlation analysis could be performed. To do this, the scores obtained were converted to rankings on each characteristic, with ties permitted.

---

4 The communality is the sum of the squared factor loadings for all factors for a given variable and gives the variance in that variable which is accounted for by all the factors. The communality measures the percentage of variance in a given variable explained by all the factors jointly, and gives a measure of the reliability of the model.
<table>
<thead>
<tr>
<th>Framework</th>
<th>Cross-disciplinary</th>
<th>Longitudinal</th>
<th>Comparability</th>
<th>Multi-disciplinary</th>
<th>Formative</th>
<th>Comprehensive</th>
<th>Flexible</th>
<th>Participant burden</th>
<th>Central burden</th>
<th>Participant expertise</th>
<th>Central expertise</th>
<th>Level of implementation</th>
<th>Scalability</th>
<th>Ownership</th>
<th>Frequency</th>
<th>Initial burden</th>
<th>Free from judgement</th>
<th>Quantitative</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIAMPI</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>REF</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIHR</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERA</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>CAHS</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>STAR METRICS</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>DIUS</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>ERIC</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>EU</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>LUMC</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>MORIA</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>PART</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>RAISS</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>RQF</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Vinnova</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>SEP</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>AERES</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CDMRP</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>NIAD-UE</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>NRF KM&amp;E</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>PBRF</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>PNECU</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>
A Spearman rank correlation analysis was performed on the rankings, excluding Level 2 of STAR METRICS, for which available data were limited. The results are shown in Table A5. Conditional formatting is used to simplify interpretation, where coloured squares have the following values: dark green, 0.7 to 1; light green, 0.5 to 0.7; yellow, −0.5 to −0.7; and red, −0.7 to −1.

The strongest correlations identified (with a correlation coefficient 0.7 or over, or -0.7 or under) are between:

- longitudinal and frequency
- longitudinal and free from judgement
- longitudinal and quantitative
- longitudinal and transparency
- frequency and quantitative
- quantitative and free from judgement
- free from judgement and transparency
- quantitative and transparency
- frequency and low central burden
- formative and flexible
- participant burden and expertise
- level of implementation and ownership
- scalability and low central burden.

These provide some initial indicators of the kind of relationships we might expect to see emerging in the following analysis.

**Factor analysis**

The next stage was to perform a factor analysis. This is a statistical approach used to describe the variability among a set of variables, which may be correlated using a smaller group of uncorrelated variables. Often it is used to reduce the number of variables in a data set by combining groups of correlated variables into a few ‘underlying’ variables. It can also be used to understand underlying factors and trends in a dataset, but here we are attempt-
Although there is no clearly identifiable elbow in the curve, both parallel analysis and optimal coordinates suggest seven factors should be sufficient to fit the data. The factanal function in statistical programme R, which conducts a factor analysis, was used to analyse the data, using varimax rotation. A test of the hypothesis that seven factors are sufficient to model the data produced a p-value of 0.0135, suggesting that the number of factors is not sufficient. This may reflect the fact that the data set is small, so

Table A5
The level of correlation between characteristics of research evaluation frameworks

<table>
<thead>
<tr>
<th></th>
<th>Cross-disciplinary</th>
<th>Longitudinal</th>
<th>Comparability</th>
<th>Multi-disciplinary</th>
<th>Formative</th>
<th>Comprehensive</th>
<th>Flexible</th>
<th>Participant burden</th>
<th>Central burden</th>
<th>Participant expertise</th>
<th>Central expertise</th>
<th>Level of implementation</th>
<th>Scalability</th>
<th>Ownership</th>
<th>Frequency</th>
<th>Initial burden</th>
<th>Free from judgement</th>
<th>Quantitative</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-disciplinary</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<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>Longitudinal</td>
<td>-0.60</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<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>Comparability</td>
<td>-0.44</td>
<td></td>
<td>0.50</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<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>Multi-disciplinary</td>
<td>0.35</td>
<td>-0.35</td>
<td>-0.26</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<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>Formative</td>
<td>0.68</td>
<td>-0.69</td>
<td>-0.66</td>
<td>0.12</td>
<td>1.00</td>
<td></td>
<td></td>
<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>Comprehensive</td>
<td>0.43</td>
<td>-0.33</td>
<td>-0.39</td>
<td>-0.07</td>
<td>0.67</td>
<td>1.00</td>
<td></td>
<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>Flexible</td>
<td>0.42</td>
<td>-0.56</td>
<td>-0.62</td>
<td>0.02</td>
<td>0.77</td>
<td>0.69</td>
<td>1.00</td>
<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>Participant burden</td>
<td>0.39</td>
<td>-0.47</td>
<td>-0.13</td>
<td>0.37</td>
<td>0.49</td>
<td>0.38</td>
<td>0.28</td>
<td>1.00</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>Central burden</td>
<td>0.23</td>
<td>-0.49</td>
<td>-0.18</td>
<td>0.10</td>
<td>0.19</td>
<td>0.02</td>
<td>0.33</td>
<td>0.40</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant expertise</td>
<td>0.34</td>
<td>-0.42</td>
<td>-0.15</td>
<td>0.40</td>
<td>0.44</td>
<td>0.43</td>
<td>0.40</td>
<td>0.86</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central expertise</td>
<td>0.52</td>
<td>-0.34</td>
<td>-0.53</td>
<td>0.37</td>
<td>0.38</td>
<td>0.23</td>
<td>0.50</td>
<td>0.24</td>
<td>0.53</td>
<td>0.30</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent of implementation</td>
<td>-0.11</td>
<td>0.08</td>
<td>-0.03</td>
<td>0.07</td>
<td>0.02</td>
<td>-0.16</td>
<td>-0.19</td>
<td>-0.03</td>
<td>-0.22</td>
<td>-0.28</td>
<td>-0.08</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>-0.34</td>
<td>0.66</td>
<td>0.31</td>
<td>0.01</td>
<td>-0.42</td>
<td>-0.11</td>
<td>-0.44</td>
<td>-0.16</td>
<td>-0.70</td>
<td>-0.33</td>
<td>-0.33</td>
<td>0.50</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ownership</td>
<td>-0.04</td>
<td>0.27</td>
<td>0.11</td>
<td>0.33</td>
<td>-0.27</td>
<td>-0.26</td>
<td>-0.29</td>
<td>-0.09</td>
<td>-0.17</td>
<td>-0.14</td>
<td>0.03</td>
<td>0.71</td>
<td>0.59</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>-0.40</td>
<td>0.76</td>
<td>-0.39</td>
<td>-0.28</td>
<td>-0.47</td>
<td>-0.28</td>
<td>-0.52</td>
<td>-0.51</td>
<td>0.71</td>
<td>-0.45</td>
<td>-0.50</td>
<td>0.21</td>
<td>0.55</td>
<td>0.26</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial burden</td>
<td>0.04</td>
<td>0.49</td>
<td>-0.01</td>
<td>-0.38</td>
<td>-0.14</td>
<td>0.27</td>
<td>0.08</td>
<td>-0.35</td>
<td>-0.13</td>
<td>-0.16</td>
<td>0.23</td>
<td>-0.38</td>
<td>0.06</td>
<td>-0.21</td>
<td>0.28</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free from judgement</td>
<td>-0.43</td>
<td>0.79</td>
<td>0.57</td>
<td>-0.57</td>
<td>-0.60</td>
<td>-0.27</td>
<td>-0.46</td>
<td>-0.59</td>
<td>-0.41</td>
<td>-0.58</td>
<td>-0.34</td>
<td>0.11</td>
<td>0.43</td>
<td>0.25</td>
<td>0.68</td>
<td>0.52</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>-0.51</td>
<td>0.79</td>
<td>0.52</td>
<td>-0.42</td>
<td>-0.68</td>
<td>-0.33</td>
<td>-0.56</td>
<td>-0.47</td>
<td>-0.36</td>
<td>-0.45</td>
<td>-0.34</td>
<td>0.19</td>
<td>0.45</td>
<td>0.28</td>
<td>0.75</td>
<td>0.44</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>-0.35</td>
<td>0.70</td>
<td>0.63</td>
<td>-0.52</td>
<td>-0.58</td>
<td>-0.27</td>
<td>-0.39</td>
<td>-0.45</td>
<td>-0.12</td>
<td>-0.30</td>
<td>-0.24</td>
<td>0.02</td>
<td>0.30</td>
<td>0.22</td>
<td>0.53</td>
<td>0.47</td>
<td>0.78</td>
<td>0.72</td>
<td>1.00</td>
</tr>
</tbody>
</table>

ing to investigate the relationship and correlation between our variables, identifying groups of characteristics that are positively and negatively correlated. Due to the non-linearity of the data, the factor analysis was performed on rankings rather than the original scores. To do this, the scores obtained were converted to rankings on each characteristic, with ties permitted. A scree plot was generated to provide an initial indication of the number of factors to include in the analysis, which is shown in Figure A2.
the interpretation of findings here should be taken in a broad qualitative sense and indeed the factors generated will not be used for further manipulation of the data set. Trial and error finds that eight factors are still not sufficient to fit the data, but that nine factors are sufficient, producing a p-value of 0.0654. Table A6 shows the results of the analysis, listing the loadings on the nine factors. Again, Table A6 is conditionally formatted so those cells containing a value above 0.6 are coloured green, and those with a number below −0.6 are coloured red.

We can look at these results to see if the small sample size may have limited the accuracy of this analysis. First, we find that the average communality is 0.85, with a lowest value of 0.73. This is high and suggests that the factors are sufficiently repre-

Table A6
Loadings of the characteristics against each of the factors

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Factor 8</th>
<th>Factor 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-disciplinary</td>
<td>−0.16</td>
<td>0.35</td>
<td>0.15</td>
<td>0.15</td>
<td>0.18</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longitudinal</td>
<td>0.70</td>
<td>−0.30</td>
<td>−0.32</td>
<td>0.12</td>
<td>−0.13</td>
<td>−0.33</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparability</td>
<td>0.25</td>
<td>−0.71</td>
<td>−0.19</td>
<td>0.15</td>
<td>−0.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-disciplinary</td>
<td>−0.38</td>
<td>0.16</td>
<td>0.22</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formative</td>
<td>−0.38</td>
<td>0.78</td>
<td>−0.32</td>
<td>0.12</td>
<td>−0.13</td>
<td>−0.33</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehensive</td>
<td>0.73</td>
<td>−0.15</td>
<td>0.33</td>
<td>0.18</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexible</td>
<td>−0.18</td>
<td>0.83</td>
<td>0.26</td>
<td>−0.13</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant burden</td>
<td>−0.32</td>
<td>0.12</td>
<td>0.18</td>
<td>0.89</td>
<td>0.16</td>
<td>−0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central burden</td>
<td>−0.16</td>
<td>0.94</td>
<td>−0.11</td>
<td>0.22</td>
<td></td>
<td>−0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant expertise</td>
<td>−0.16</td>
<td>0.23</td>
<td>0.20</td>
<td>−0.17</td>
<td>0.87</td>
<td>0.23</td>
<td>−0.11</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Central expertise</td>
<td>0.42</td>
<td>0.59</td>
<td>0.38</td>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent of implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−0.12</td>
<td>0.90</td>
<td>−0.11</td>
<td></td>
<td>−0.12</td>
</tr>
<tr>
<td>Scalability</td>
<td>0.27</td>
<td>−0.17</td>
<td>−0.55</td>
<td>0.51</td>
<td></td>
<td>−0.15</td>
<td>0.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ownership</td>
<td>0.15</td>
<td>−0.19</td>
<td>0.86</td>
<td>0.32</td>
<td></td>
<td></td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>0.52</td>
<td>−0.25</td>
<td>−0.62</td>
<td>0.17</td>
<td>−0.20</td>
<td>−0.11</td>
<td>−0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial burden</td>
<td>0.84</td>
<td>0.26</td>
<td>−0.37</td>
<td>−0.16</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free from judgement</td>
<td>0.76</td>
<td>−0.32</td>
<td>−0.23</td>
<td>0.15</td>
<td>−0.29</td>
<td>−0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>0.76</td>
<td>−0.41</td>
<td>−0.21</td>
<td>0.20</td>
<td>−0.14</td>
<td>−0.15</td>
<td>−0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>0.70</td>
<td>−0.34</td>
<td>−0.11</td>
<td>−0.31</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure A2
Scree plot to determine the number of factors used in the analysis of the characteristics of frameworks

![Scree plot to determine the number of factors used in the analysis of the characteristics of frameworks](image-url)
sentative of the data set. Second, we find that the first five factors have a high loading on several variables. This is less clear for the final four factors, and on that basis we will focus our analysis on only those five factors. Third, mapping these loadings against the frameworks we do not find that any particular framework(s) are dominating the factors produced.

We can thus identify the key characteristics associated with each factor, using a cut-off of 0.6. There is no particular causation indicated numerically in the relationship between factors; however, we can draw on our experience with research evaluation to suggest the relationship between the groups of characteristics related through each factor, and thus turn the combination between characteristics into a series of statements as presented below, in order of significance:

- Factor 1: quantitative approaches can produce longitudinal data and are free from judgement and transparent but have a high initial burden. This factor is the strongest relationship found, accounting for 19.8% of the variance in the dataset.
- Factor 2: formative approaches tend to be comprehensive and flexible but not comparable (17.7% of variance).
- Factor 3: approaches that have a high central burden tend to be applied infrequently (12.3% of variance).
- Factor 4: a high level of implementation is associated with ownership (11.4% of variance).
- Factor 5: high participant burden is associated with a high level of participant expertise required (10.9% of variance).

The last four factors are not significant (accounting for less than 10% of variance).

We can also then map these loadings against the scores for each of the six frameworks investigated in detail to establish the extent to which each framework is aligned with the factors identified, to provide some illustrative examples of the relationships indicated. We find that the following frameworks are associated with the factors:

- Factor 1: negatively associated with Productive Interactions, REF and CAHS
- Factor 2: associated with Productive Interactions and CAHS; negatively associated with ERA and STAR METRICS Level 1
- Factor 3: associated with Productive Interactions and CAHS; negatively associated with NIHR Dashboard
- Factor 4: negatively associated with Productive Interactions and CAHS
- Factor 5: negatively associated with NIHR Dashboard, STAR METRICS Level 1 and ERA.

We can also plot the relationships between the five factors and the six frameworks, as shown in Figure A3.

**Analysis of tools**

The tools were also scored against the same characteristics, although ‘ownership’ and ‘extent of implementation’ were excluded, since these characteristics are less relevant in the context of research evaluation tools. The scores given to each of the tools are presented in Table A7.

These scores were converted into rankings, and a correlation analysis performed producing the correlation matrix shown in Table A8, which indicates the interaction between the different characteristics in relation to research evaluation tools. Again, conditional formatting is used to simplify interpretation, where coloured squares have the following values: dark green, 0.7 to 1; light green, 0.5 to 0.7; yellow, −0.5 to −0.7; and red, −0.7 to −1.

It is not possible to conduct a factor analysis on these data since there are a greater number of characteristics than tools. However, we can draw some conclusions from the analysis of this correlation matrix. We identify two groups of characteristics which are positively correlated with each other and negatively correlated with the characteristics in the other group:

- Cross-disciplinary, formative, flexible and multidisciplinary are associated. Comprehensive is also associated with this group to a lesser extent. These characteristics are negatively associated with the characteristics listed in Group 2.
Figure A3
Plot showing the alignment of the six frameworks investigated in detail with the five most significant factors

Table A7
Scoring of research evaluation tools against characteristics

<table>
<thead>
<tr>
<th></th>
<th>Cross-disciplinary</th>
<th>Longitudinal</th>
<th>Comparability</th>
<th>Multi-disciplinary</th>
<th>Formative</th>
<th>Comprehensive</th>
<th>Flexible</th>
<th>Participant burden</th>
<th>Central burden</th>
<th>Participant expertise</th>
<th>Central expertise</th>
<th>Level of implementation</th>
<th>Scalability</th>
<th>Ownership</th>
<th>Frequency</th>
<th>Initial burden</th>
<th>Freedom from judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliometrics</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Surveys</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Logic models</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Case studies</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Economic analysis</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Peer review</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Data mining</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Interviews</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Document review</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Site visits</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Taking these two groups, we can map the alignment of tools investigated with Group 1 and Group 2 characteristics by summing their scores on each of the characteristics strongly associated with the group, and dividing by the number of characteristics in the group. This creates a normalised Group 1 and Group 2 score for each tool. These scores are plotted on pages 9–12.

**Table A8**

Correlations between characteristics of research evaluation tools

<table>
<thead>
<tr>
<th></th>
<th>Cross-disciplinary</th>
<th>Longitudinal</th>
<th>Comparability</th>
<th>Multi-disciplinary</th>
<th>Formative</th>
<th>Comprehensive</th>
<th>Flexible</th>
<th>Participant burden</th>
<th>Central burden</th>
<th>Participant expertise</th>
<th>Central expertise</th>
<th>Scality</th>
<th>Frequency</th>
<th>Initial burden</th>
<th>Free from judgement</th>
<th>Quantitative</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-disciplinary</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<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>Longitudinal</td>
<td>-0.90</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<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>Comparability</td>
<td>-0.73</td>
<td>0.80</td>
<td>1.00</td>
<td></td>
<td></td>
<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>Multi-disciplinary</td>
<td>0.57</td>
<td>-0.95</td>
<td>-0.78</td>
<td>1.00</td>
<td></td>
<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>Formative</td>
<td>0.80</td>
<td>-0.60</td>
<td>-0.50</td>
<td>0.69</td>
<td>1.00</td>
<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>Comprehensive</td>
<td>0.76</td>
<td>-0.63</td>
<td>-0.61</td>
<td>0.71</td>
<td>0.75</td>
<td>1.00</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>Flexible</td>
<td>0.94</td>
<td>-0.92</td>
<td>-0.71</td>
<td>0.92</td>
<td>0.76</td>
<td>0.68</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant burden</td>
<td>0.60</td>
<td>-0.58</td>
<td>-0.33</td>
<td>0.51</td>
<td>0.38</td>
<td>0.61</td>
<td>0.55</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central burden</td>
<td>0.52</td>
<td>-0.50</td>
<td>-0.05</td>
<td>0.49</td>
<td>0.24</td>
<td>0.34</td>
<td>0.61</td>
<td>0.58</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant expertise</td>
<td>0.41</td>
<td>-0.24</td>
<td>-0.23</td>
<td>0.23</td>
<td>0.73</td>
<td>0.47</td>
<td>0.41</td>
<td>0.42</td>
<td>0.10</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central expertise</td>
<td>-0.40</td>
<td>0.43</td>
<td>0.74</td>
<td>-0.47</td>
<td>-0.38</td>
<td>-0.57</td>
<td>-0.26</td>
<td>-0.07</td>
<td>0.38</td>
<td>-0.22</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>-0.92</td>
<td>0.84</td>
<td>0.74</td>
<td>-0.86</td>
<td>-0.80</td>
<td>-0.70</td>
<td>-0.93</td>
<td>-0.61</td>
<td>-0.49</td>
<td>-0.63</td>
<td>0.39</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>-0.81</td>
<td>0.84</td>
<td>0.59</td>
<td>-0.78</td>
<td>-0.46</td>
<td>-0.45</td>
<td>-0.78</td>
<td>-0.80</td>
<td>-0.58</td>
<td>-0.33</td>
<td>0.15</td>
<td>0.81</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial burden</td>
<td>-0.66</td>
<td>0.86</td>
<td>0.75</td>
<td>-0.75</td>
<td>-0.23</td>
<td>-0.33</td>
<td>-0.62</td>
<td>-0.51</td>
<td>-0.25</td>
<td>0.00</td>
<td>0.44</td>
<td>0.59</td>
<td>0.83</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free from judgement</td>
<td>-0.83</td>
<td>0.88</td>
<td>0.76</td>
<td>-0.89</td>
<td>-0.61</td>
<td>-0.74</td>
<td>-0.82</td>
<td>-0.46</td>
<td>-0.41</td>
<td>-0.32</td>
<td>0.64</td>
<td>0.81</td>
<td>0.64</td>
<td>0.68</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>-0.94</td>
<td>0.88</td>
<td>0.73</td>
<td>-0.96</td>
<td>-0.83</td>
<td>-0.74</td>
<td>-0.90</td>
<td>-0.38</td>
<td>-0.37</td>
<td>-0.36</td>
<td>0.54</td>
<td>0.85</td>
<td>0.63</td>
<td>0.61</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>-0.86</td>
<td>0.94</td>
<td>0.91</td>
<td>-0.90</td>
<td>-0.65</td>
<td>-0.69</td>
<td>-0.89</td>
<td>-0.42</td>
<td>-0.33</td>
<td>-0.30</td>
<td>0.60</td>
<td>0.83</td>
<td>0.68</td>
<td>0.77</td>
<td>0.89</td>
<td>0.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- Longitudinal, scalability, frequency, free from judgement, quantitative and transparent are associated. Comparability and initial burden are also associated with this group to a lesser extent. These characteristics are negatively associated with characteristics in Group 1.

We can also note that frequency is associated with low participant burden.
Appendix C  Finding out more: international examples of research evaluation frameworks

In this section, we provide an overview of eight research evaluation frameworks and more detailed analysis of a further six frameworks, selected in collaboration with AAMC:

• the Canadian Academy of Health Science Payback Framework (Canada)
• Excellence in Research for Australia (ERA) (Australia)
• the National Institute of Health Research Dashboard (England)
• the Research Excellence Framework (REF) (UK)
• Productive Interactions (Netherlands and European Commission).
• Science and Technology for America’s Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science (STAR METRICS) (US)

These frameworks were selected on the basis of relevance and interest to AAMC, and consist of a mix of national evaluation systems and health research specific approaches. Reviews have been conducted, drawing on documentary review and key informant interviews. These are reported using a standard reporting template, covering the following areas:

• origin and rationale: the factors that led to the development of the system, including its aims, previous evaluation systems, the theoretical and conceptual basis of the framework, and the wider context

• scope: a range of factors covering the scope of the approach, including unit of analysis, types of impact covered, the extent of application and the timescale involved

• measurement: data collection approaches and methodologies, and the way in which data are aggregated and analysed

• application to date: current status of application, including the uses of the data, the scale of application, the main audience for the data, and any compromises that have been made in the approach

• analysis: our assessment of the effectiveness of the approach, including whether it achieves its aims and its strengths and weaknesses; this section draws on RAND Europe expertise as well as more formal documentary evidence

• wider applicability: an assessment of the applicability of this approach more widely, including transferability, its critical reception, costs of implementation and transaction costs.

The level of content in these areas differs between the frameworks, depending on the availability of documentary evidence, perhaps also reflecting the differing levels of development of the frameworks investigated.

In addition to the six frameworks explored in detail we also present short, two-page summaries of a further eight research evaluation frameworks. Nineteen frameworks were identified, drawing on internal knowledge and documentary evidence, as potential candidates for inclusion in this format. Frameworks from this long list were then prioritised for inclusion on the basis of their geographical coverage and coverage in previous studies by RAND Europe (Brutscher, Wooding and Grant, 2008; and Grant et al., 2009). Research frameworks not analysed elsewhere were included, where possible,
to broaden the range of approaches investigated. A broad geographical coverage was also sought, to facilitate understanding and learning from best practice internationally. These shorter reviews are not constrained to a common structure; instead they are written to highlight the points of interest and novelty from each of the approaches described. The aim is to provide an illustration of the range of approaches used in different contexts internationally, and to feed into the wider analysis outlined in this report. These are the eight frameworks we have described in this way:

- the Evaluation Agency for Research and Higher Education (AERES) (France)
- the Congressionally Directed Medical Research Programme (CDMRP) (US)
- the National Commission for University Evaluation and Accreditation (CONEAU) (Argentina)
- the National Institute for Academic Degrees and University Evaluation (NIAD-UE) (Japan)
- the Knowledge Management and Evaluation Directorate of the National Research Fund (NRF KM&E) (South Africa)
- the Performance Based Research Fund (PBRF) (New Zealand)
- the Spanish State Programme for University Quality Assurance (PNECU) (Spain)

The reviews of research evaluation frameworks are presented below, starting with the more detailed descriptions of six frameworks, as described above, then followed by the brief summaries of the further eight frameworks.
The Canadian Academy of Health Sciences framework

**Brief description**

The Canadian Academy of Health Sciences (CAHS) framework is an approach to measuring return on investment in health research. It is designed to track impacts from research through translation to end use. It also demonstrates how research influences feedback upstream and the potential effect on future research. The framework tracks impacts under the following categories of health research:

- advancing knowledge
- capacity building
- informing policies and product development
- health and health sector benefits
- broader socio-economic benefits.

The CAHS framework captures outcomes of interest for different audiences. It is intended to be applicable in the evaluation of health research impacts by all funders of health research in Canada. There are six categories of health research funders in Canada: federal, provincial, industry, higher education sector, private non-for-profit sector and foreign investment. The approach aims to provide breadth, depth and flexibility to achieve:

- usefulness to a full range of funders and research types
- compatibility with existing structures in Canada
- transferability to international comparison
- the ability to identify the full spectrum of potential impacts.

**Origin and rationale**

This measurement tool was developed by the Canadian Academy of Health Sciences in 2009 to provide an organised framework for capturing returns on investment in health research. The CAHS framework is an adapted ‘payback model’ (Buxton and Hanney, 1996, 1998, 2005; Buxton, Hanney and Jones, 2004; Wooding et al., 2005) that aims to capture specific impacts in multiple domains, at multiple levels and for a wide range of audience. A set of indicators and metrics for use in this framework was developed to evaluate the returns on investment in health research. Figure A4 shows the CAHS Logic Model Framework built on a payback logic model, showing the CAHS logic model (at the top) and how it compares to the payback framework (at the bottom).

As explained above, impacts are classified into five categories: advancing knowledge, capacity building, informing policies and product development, health and health sector benefits, and broader socio-economic benefits. The CAHS payback model combines an impact category (each one of the five impact categories are represented in the top half of the diagram) approach with a logic model (bottom half of the diagram).

---

5 These are based on the categorisation provided by Statistics Canada.
6 The Canadian Academy of Health Sciences was established in 2004. It is one of the three Canadian academies of science whose mission is to provide credible independent expert assessments on urgent health issues. It differs from the Canadian Institute of Academic Medicine in that it uses a multi-disciplinary approach to provide a holistic perspective on health related matters.
7 A set of 66 preferred indicators were identified. These are summarised in the section ‘Measurement’, below.
8 Figure A6 illustrates how the five impact categories are represented in the top half of the CAHS logic model framework.
The health system in Canada is funded by multiple stakeholders (provincial, federal, high education, industry and not-for-profit organisations) and impacts from research are wide ranging. With significant increase in investments in health research there have been growing expectations for returns and a need to understand societal impacts. The CAHS identifies a number of factors that have led to the need to develop a robust method and indicators for capturing impacts of health research including (CAHS, 2007):

- lack of public understanding of the value of research and its applicability to current issues in health care at a time of unsurpassed concern about accessible, affordable, high-quality health care in a publicly funded system
- failure to adequately measure the benefits of fundamental and applied health research and to properly convey them in a meaningful fashion to policymakers and the public
- an increasingly common view that health care (and by association, health research) is a cost driver consuming an ever greater share of provincial resources at the expense of other sectors
- growing general concern about expenditure accountability in the aftermath of widely publicised instances of misuse in both the public and private sectors in Canada and abroad
The framework aims to contribute to the social and economic well-being of the country through improvement in health. However, tracking impacts from research through translation to end use and understanding how these feed back to upstream activity and their potential influence on future research has an international perspective: investment in health research is also aimed at contributing to developing the country’s competitiveness and keeping it at the cutting edge of science and technology through collaboration (Armstrong et al., 2005).

Scope

The impact framework was developed to capture research that is relevant to health based on four identified pillars. These pillars, defined by the Canadian Institute of Health Research (CIHR), provide the scope for assessment (CIHR, 2007, n.d.):

- **biomedical**: research with the goal of understanding normal and abnormal human functioning at the molecular, cellular, organ system and whole body levels, including development of tools and techniques to be applied for this purpose; developing new therapies or devices that improve health or the quality of life of individuals, up to the point where they are tested on human subjects; studies on human subjects that do not have a diagnostic or therapeutic orientation
- **clinical**: research with the goal of improving the diagnosis and treatment (including rehabilitation and palliation) of disease and injury, and improving the health and quality of life of individuals as they pass through normal life stages; research on, or for the treatment of, patients
- **health services**: research with the goal of improving the efficiency and effectiveness of health professionals and the health care system, through changes to practice and policy; health

The CAHS framework seeks to provide a common approach to health research funders in tracking health research impacts. Before the development of this assessment tool, there was no joined up evaluation approach in the Canadian health research landscape although most funders had put in place evaluation frameworks, most of which were based on a logic model. For example, the Canadian Institute of Health Research had adopted the payback model and defined the five impact categories used in the Canadian Academy of Health Sciences framework. However, other frameworks, such as the integrated Results Based Management and Accountability Framework and Risk-based Audit Framework (Canadian Research Chairs), and the Performance, Audit and Evaluation Strategy (Genome Canada), were also in use. The CAHS payback model combines an impact category approach with a logic model and provides a standardised research evaluation approach that has two main objectives:

- to enable comparison of evaluations
- to enable identification of the full spectrum of potential impacts including unexpected outcomes of research.
services research is a multi-disciplinary field of scientific investigation that studies how social factors, financing systems, organisational structures and processes, health technologies and personal behaviours affect access to health care, the quality and cost of health care and, ultimately, Canadians’ health and well-being.

- **social, cultural, environmental and population health**: research with the goal of improving the health of the Canadian population, or of defined sub-populations, through a better understanding of the ways in which social, cultural, environmental, occupational and economic factors determine health status.

These four areas generate a wide range of health-related outputs. Understanding their returns on investment in research requires tracking the impact of these outputs and unravelling the causal pathways that lead to the impacts so as to facilitate intervention. The CAHS framework tracks health impacts of these research areas based on the five categories previously mentioned (advancing knowledge, capacity building, informing policies and product development, health and health sector benefits, and broader socio-economic benefits).

**Measurement**

The CAHS framework uses a set of preferred indicators and metrics defined for each of the five impact categories, which are listed in Table A9.

The CAHS framework links impact categories to a logic model and in so doing provides a matrix for collecting data in a coordinated and replicable manner. It also helps to identify where health impacts are most likely to occur.

The logic model builds in each one of the four pillars – basic biomedical research, applied clinical research, health services and policy research, and population and public health research – and aims to understand how each one of the pillars progresses to impacts. Then impact categories are cross mapped for each pillar to create specialised frameworks. Figure A6 illustrates impact categories for biomedical research.

The five impact categories are colour-coded in the CAHS payback model (in the top half of the diagram). The colour-coded shading indicates where the different impacts are most likely to occur. In biomedical research there are primary outputs resulting from knowledge advancement (research results) as well as secondary outputs in industry, and contributions to the research agenda, and to public information groups. Capacity building occurs in the research, both in academia and industry, while outputs feeding into informing the decisionmaking process occur at the level of the industry, government, the research community, and the public and the health care system. Health benefits accrue at the level of determinants of health and contributed to improved health and well-being of the population. Economic and social impacts occur because of the activity. Finally, health research and research commercialisation by industry contribute to improvements in the health of the population, and in so doing to economic and social prosperity.

**Application to date**

The CAHS framework is used to assess returns on investment in health research. In so doing, it is intended to address a number of factors, including accountability, measurement of the benefits of health research, appropriateness of funded research, supporting policymaking processes and so on.

---

9 See CAHS (2011) for details of a set of 66 preferred indicators.
10 Impacts may nevertheless occur outside the shaded areas. An evaluation based on CAHS framework is encourage to consider the possibility of impacts occurring in areas where they are least expected.
11 As noted previously the shading is an indication of where impacts may most likely be found. The shading is not intended to exclude the possibility of impacts occurring elsewhere.
Table A9
Preferred indicators and metrics for impact categories of the CAHS framework

<table>
<thead>
<tr>
<th>Impact category</th>
<th>Indicator type</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advancing knowledge includes new discoveries and breakthroughs from health research and contributions to scientific literature</td>
<td>Quality indicators</td>
<td>Relative citation impact, highly cited publications, publications in high-quality outlets</td>
</tr>
<tr>
<td></td>
<td>Activity indicators</td>
<td>Share of publications, publication counts</td>
</tr>
<tr>
<td></td>
<td>Outreach</td>
<td>Co-author analysis, field analysis of citations</td>
</tr>
<tr>
<td></td>
<td>Contextual and cultural</td>
<td>Relative activity index</td>
</tr>
<tr>
<td></td>
<td>Aspiration</td>
<td>Expanded relative citation impact, relative download rate, research diffusion</td>
</tr>
<tr>
<td>Capacity building includes development and enhancement of research skills for individuals and teams</td>
<td>Personnel</td>
<td>Graduated research students in health-related subjects; numbers of research and research-related staff in Canada</td>
</tr>
<tr>
<td></td>
<td>Funding</td>
<td>Levels of additional research funding</td>
</tr>
<tr>
<td></td>
<td>Infrastructure</td>
<td>Infrastructure grants, percentage of activity grants with infrastructure support</td>
</tr>
<tr>
<td></td>
<td>Aspiration indicators</td>
<td>Receptor capacity, absorptive capacity</td>
</tr>
<tr>
<td>Informing decisionmaking includes the impacts of research in areas of science, public, clinical and managerial decisionmaking practice policy</td>
<td>Health related (health care, public health, social care and other health-related systems)</td>
<td>Use of research in guidelines, survey of public health policymakers, research reported use of findings outside health, research cited in ongoing health professional education material</td>
</tr>
<tr>
<td></td>
<td>Research</td>
<td>Citations analysis of successful funding applications, consulting to policy, requests for research to support policy, research used in curricula for new researchers</td>
</tr>
<tr>
<td></td>
<td>Health products industry</td>
<td>Number of patents licensed, clustering and co-location, consulting to industry; collaboration with industry, use of research in stage reports by industry</td>
</tr>
<tr>
<td></td>
<td>General public</td>
<td>Research cited in advocacy publications, public lectures given</td>
</tr>
<tr>
<td></td>
<td>Aspirational indicators</td>
<td>Media citation analysis, citation in public policy documents</td>
</tr>
<tr>
<td>Health impacts include advances in prevention, diagnosis, treatment and palliation when related to research</td>
<td>Health status</td>
<td>Prevalence, incidence, potential years life lost, quality-adjusted life years (QALYs), patent reported outcome measures</td>
</tr>
<tr>
<td></td>
<td>Determinants of health</td>
<td>Examples include obesity, alcohol consumption, waiting times, adherence to clinical guidelines etc.</td>
</tr>
<tr>
<td>Broader economic and social impacts</td>
<td>Activity impact</td>
<td>Economic rent (labour rent) – benefit accrued through the action of research as opposed to outputs of research – the economic benefit of employing people in health research rather than in another capacity</td>
</tr>
<tr>
<td></td>
<td>Commercialisation</td>
<td>Licensing returns, product sales revenues, valuation of spin out companies, producer rent and spillover effect – economic benefit accrued through sales and revenues of commercialised research findings</td>
</tr>
<tr>
<td></td>
<td>Health benefit</td>
<td>Health benefit in QALYs per health care dollar, health benefit in patient reported outcome measures per health care dollar – the net benefit of improving health</td>
</tr>
<tr>
<td></td>
<td>Well-being</td>
<td>Happiness, level of social isolation</td>
</tr>
<tr>
<td></td>
<td>Social benefits</td>
<td>Social economic status – identifying whether changes in socio-economic status correlate with research impacts (assessed through specific research studies)</td>
</tr>
</tbody>
</table>
The framework is intended to be applicable in the evaluation of health research impacts by all funders of health research in Canada – federal, provincial, industry, higher education sector, private non-for-profit sector and foreign investment. It is designed for application at various levels: individual, institutional, provincial, national or international. Although its application has so far been at the institutional or organisational level, it is also intended to provide a national perspective that would allow comparability at the international level.

The main users so far have been public funders of health research – in particular CIHR at the federal level, and the National Alliance of Province Health Research Organizations (NAPHRO) at the provincial level. There have also been specific users at provincial level such as Alberta Innovates. Uptake is not mandatory, but application by CIHR, the main health research funder in Canada, is likely to encourage broader uptake.

Flexibility is one of the key aspects of the framework as it is designed for application by all funders of health research and to describe impacts at different levels. It is therefore a ‘skeleton’ that needs to be tailored for different evaluations. The main alterations are generally in the choice of indicator groups, although
there is also scope for modification of the framework to better fit with organisational needs.

**Analysis**

The framework provides a basis for identifying:

- where data should be collected
- what data should be collected
- how to perform evaluation of health research to ensure comparability and comprehensiveness.

The combination of breadth, depth and flexibility suggests that the framework can be useful to a wide range of funders, compatible with what is already in place in Canada, transferable to international comparisons, and able to identify many potential impacts. One of the main advantages of the CAHS framework, therefore, is that it offers a tailor-made approach to evaluation while offering a basis for comparison. Nonetheless, such a tailor-made approach also implies fairly substantial human resource investments in its application. Furthermore, it is not clear how modifications of the framework or changes in indicators resulting from emerging technological advancements for example would be managed in order to ensure that requirements such as comparability are maintained.
These are some of the caveats identified about the framework (CAHS, 2009):

- The framework can help guide evaluations but it does not provide the questions (or refine them).
- The framework cannot resolve all complexities easily or immediately. If applied prospectively, it will take time and, depending on the scope, could become costly to apply.
- It is only as good as the selection of domains, indicator sets and data within indicators to resolve attribution progressively, to achieve a balanced perspective, to avoid the halo effect. The menu of 66 indicators provided is only a start – more are required.
- Balanced perspectives may not provide clear-cut answers.

The CAHS framework is at the initial stages of its application and how well it achieves its aims can only be determined with time and wider application. However, a critical assessment of the framework suggests some strengths and weaknesses:

- Strengths: generalisable (it has wide application relevance for funders and aspects of health research); it can access unexpected outcomes through the logic model (not just a metric driven approach); and it is pan-Canadian (the development of the framework sought to reflect inputs from a wide range of stakeholders to allow comparison across funding agencies and provinces) and has therefore great potential for uptake.
- Weakness: funders are required to know how to use an evaluation framework – a skill they may not possess; and it requires funders to review their existing evaluation frameworks and consider modifications for alignment with the CAHS framework to enhance comparability with other funders.

Greater uptake of the CAHS framework by funders over time will be necessary to assess how it works in practice.

### Wider applicability

The Canadian Academy of Health Sciences framework is designed to evaluate the impacts of health research in Canada in a flexible way that would allow all health research funders to apply it. The framework therefore aims to meet the following criteria:

- usefulness to a full range of funders and research types
- compatibility with existing structures in Canada
- transferability to international comparison
- ability to identify the full spectrum of potential impacts.

Set-up costs at implementation can be high since the framework relies on standardising data collection. Transaction costs may vary (in time and money) depending on the selected indicator set. Highly quantitative indicators have low transaction but high implementation costs, while the converse is true for highly qualitative indicators. A combination of the two types of indicators is generally used for evaluations. Users who previously used evaluation frameworks based on a logic model may find it relatively easy to implement. Staff at CIHR, one of the users so far, had previously adopted a payback framework. The ease with which staff at the other two organisations NAPHRO (the collective group for the provincial health research organisations) and Alberta Innovates (a provincial health research organisation) have been able to implement the CAHS framework remains unclear.

Although it is too soon to assess feedback on the framework, it appears to have generated interest outside health research in Canada, Australia and the US, as well as in Spain, where implementation by the Catalan Agency for Health Information and Quality (CAHIAQ) is currently under way through the Social Impact of Health Research (ISOR) project. Outside health research, it has been used to inform analysis of returns on investment in health care approaches, a performance measurement approach for social policy research, and to inform thinking on value for money in the procurement of pharmaceutical products.
## SWOT analysis

**Figure A7**  
A SWOT analysis for the CAHS framework

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very comprehensive</td>
<td>Resource intensive</td>
</tr>
<tr>
<td>Flexible</td>
<td>Complicated</td>
</tr>
<tr>
<td>Developed through engagement, and has strong buy-in</td>
<td>Not easily comparable</td>
</tr>
<tr>
<td>Formative</td>
<td>Implementation challenging</td>
</tr>
<tr>
<td>Looks at process as well as outputs and impacts</td>
<td>Developed by committee</td>
</tr>
<tr>
<td>Concept of an indicator library</td>
<td>Requires participant expertise</td>
</tr>
<tr>
<td>Aligned with main funders, framework</td>
<td>Not ranking – hard to use to allocate funding</td>
</tr>
<tr>
<td></td>
<td>Large burden on participants</td>
</tr>
<tr>
<td></td>
<td>Not multi-disciplinary</td>
</tr>
<tr>
<td></td>
<td>Definitional ambiguity between outputs and outcomes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unified but flexible approach</td>
<td>No implementing owner</td>
</tr>
<tr>
<td>Potential to build an indicator platform and toolkit</td>
<td>Slow uptake</td>
</tr>
<tr>
<td>– opportunity for international uptake and wider comparability</td>
<td>Dependent on CIHR endorsement</td>
</tr>
</tbody>
</table>
Excellence in Research for Australia

Brief description

The Excellence in Research for Australia (ERA) initiative is the first scheme employed in Australia to review and assess the quality of research in Australia comprehensively. The first full scale deployment took place in 2010 after three years of consultation and development. The approach is intended to be streamlined and transparent, minimising burden and building the trust of stakeholders, and the goal is to provide an internationally benchmarked overview of the quality of research conducted at all eligible higher education institutions (HEIs) in Australia. The process is run by the Australian Research Council (ARC) for the Australian government, and though it has only filled accountability and advocacy functions to date, it is likely to be used for funding allocation in future deployments.

Glossary

UoE: The unit of evaluation is the unit of analysis in ERA. It is at the discipline level, and uses a standard classification system in Australia to define the discipline boundaries.

Low volume threshold: The minimum level of research that needs to be produced for an institution to be deemed ‘research active’ in a particular UoE and hence eligible for ERA. This is defined by number of articles, at 50 articles over the six-year research period where bibliometrics are used, or 30 articles where peer review is used.

Origin and rationale

The aims of the ERA initiative are to:

• establish an evaluation framework that gives government, industry, business and the wider community assurance of the excellence of research conducted in Australia’s HEIs
• provide a national stock-take of discipline-level research strength and areas where there is opportunity for development in Australia’s HEIs
• identify excellence across the full spectrum of research performance
• identify emerging research areas and opportunities for further development
• allow for comparisons of Australia’s research nationally and internationally for all discipline areas.

The approach is metrics-based, with the intention that burden will be kept to a minimum, and builds on the existing Composite Index, which had been used since 1995 to inform the distribution of a portion of Australian research funds. The Composite Index used a formula to calculate each university’s share of total research activity in the country, on the basis of research funding awarded, numbers of publications, and graduate degrees completed (Innovation, n.d.). This was used in differing ways in the allocation of the majority of the institutional research funding allocated in Australia, which is roughly half the country’s total funding for research (Innovation, 2008). The approach was criticised partly because of the inaccuracy of the information used to calculate the allocation (an audit conducted by KPMG in 1997 of publication lists submitted by universities found an error rate of 34 per cent) and
Appendix C

63

also because of potential perverse incentives introduced (Harman, 2000; Hicks, 2009). The approach rewarded volume of publications, rather than quality, and Butler (2003) suggested that this had led to a higher level of publication in lower quality journals, and a decrease in the relative impact of Australian research internationally. Although this conclusion was not universally accepted (see for example Redden, 2008), the criticism and also the experience in other countries, most notably in the UK where the Research Assessment Exercise (RAE) had led to increased performance in the higher education sector (Gläser et al., 2002), led to moves towards assessing quality in Australian research.

Initially, this took shape as the Research Quality Framework, effectively a modified research assessment, which was developed by the Howard government. The RQF was widely unpopular, particularly among the research intensive institutions in Australia, and was also roundly criticised in the run up to elections in 2007 by Senator Kim Carr, the Shadow Minister for Industry, Innovation, Science and Research. Following the election, the new Rudd government announced ERA, supposedly to provide a ‘fairer deal’ for Australian researchers. The ERA approach was subject to open consultation and piloting, but was introduced fairly rapidly, with submissions for the 2010 evaluation completed by August 2010, only two and a half years after the idea was initially floated.

The metrics-based approach is intended to be more streamlined and transparent than the RQF’s peer review system would have been, though peer review is still used in some of the arts and humanities disciplines where bibliometric methods to assess quality are not well established. In all cases, the final judgements on ratings are made by expert panels, on the basis of the range of indicators collected. In the first round in 2010, the results were not used to allocate funding, but it is envisaged that in the long term ERA will be used to comparatively rank universities, and will have funding implications.

Scope

ERA is intended to provide an overview of the quality and performance of research across all Australian HEIs, so it is compulsory for all HEIs to participate, subject to eligibility. The first full deployment was conducted in 2010, and the second is scheduled for 2012, so it is reasonable to assume that it will continue to be used bi-annually.

The unit of analysis for ERA is termed a ‘unit of evaluation’ (UoE). This broadly corresponds to individual research disciplines at each institution, and is defined using Australian and New Zealand Standard Research Classification Field of Research (FoR) codes. These are one of three standard research classification systems across Australia and New Zealand, which provide a series of classifications by FoR. The fields were updated in 2008 and include 22 divisions (two-digit codes) further subdivided into (four-digit) ‘groups’ and (six-digit) ‘fields’. For example, the ‘field’ of Meteorology (Code 040107) falls under the ‘group’ of Atmospheric sciences (Code 0401), which in turn sits within the ‘division’ of Earth Sciences (Code 04). Evaluation is conducted at the two- and four-digit code levels for the UoE, and UoEs are also grouped into eight discipline ‘clusters’: physical, chemical and earth sciences; humanities and creative arts; engineering and environmental sciences; social, behavioural and economic sciences; mathematical, information and computing sciences; biological sciences and biotechnology; biomedical and clinical health sciences; and public and allied health services.

In addition, each UoE must meet what is termed a ‘low volume threshold’. This means that institutions must reach a certain minimum level of activity in a particular discipline to be entered for assessment. In those UoEs where citation analysis is used to assess quality, the low volume threshold is 50 articles over the six-year research outputs reference period. Where peer review is used, the threshold is 30 outputs over the same period. To take into account the different means of dissemination in different fields, some outputs count more than once. For example, books count as equivalent to five journal publications for the purposes of this minimum activity assessment. Where the low volume threshold is not met, the institution is considered not to be active in that area of research, and receives the rating ‘n/a’.

ERA covers a range of different types of impact, though the focus is on quality. In the ‘research appli-
Data are collected over a fixed reference period, and this has not changed between the initial 2010 evaluation and the upcoming evaluation in 2012. Table A10 shows the time period over which the different elements are evaluated, and how this was implemented in 2010, and will be implemented in 2012.

The research submitted is assessed against a rating scale, which is intended to be consistent with that used in other countries. Table A11 shows the scale used in 2010, and this is unlikely to change in 2012.

Research is assessed on a UoE level as described in the previous section. Although assessment is largely standardised across different UoEs, there are subject-specific differences, for example, the use of peer review in place of citation indicators in the assessment of quality. Here the split is broadly between social sciences, arts and humanities and the sciences, as might be expected. Similarly, those other indicators which will be applied differ between UoEs, reflecting what is appropriate for particular disciplines. For example, for the medical sciences the applied indicators included in 2010 were research commercialisation income, citation on NHMRC guidelines, and patents, while in architecture the indicator guideline citations is replaced with registered designs, and for philosophy only data on research commercialisation income were sought.

These differences between UoEs are also present in the forthcoming 2012 process (ARC, n.d.). Where peer review is used, the committee may appoint external reviewers if there is not sufficient expertise within the committee to perform the assessment internally. Quality assessments are not reported for individual outputs, but on the overall sample submitted for review, which will be 20 per cent of all outputs. Even within that 20 per cent, reviewers were not required to read all outputs, rather a selection sufficient for a robust evaluation. In the field of the creative arts, where creative work may be submitted, a research statement accompanies the sub-

---

12 Details of those UoEs which were expected to use peer review instead of citations and vice versa, along with other differences in the indicators used in the evaluation process, are provided in the ERA 2010 disciplines matrices available at http://arc.gov.au/era/era_2010/archive/key_docs10.htm.
kind of aggregated outputs that are produced are shown in figures A8 and A9.

**Application to date**

The ERA approach was trialled in 2009 for two discipline clusters – physical and chemical sciences, and humanities and creative arts. The approach was broadly found to be a success and no significant changes were made to the overall principles of the approach. However, there were some minor changes between the pilot and full roll-out in 2010:

- Researcher eligibility requirements were clarified, with an additional employment category provided to accommodate casual staff.

<table>
<thead>
<tr>
<th>Table A10</th>
<th>Time periods for different sets of indicators in ERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>Period between evaluations (years)</td>
</tr>
<tr>
<td>Staff eligibility</td>
<td>Census date</td>
</tr>
<tr>
<td>Citation data period</td>
<td>7.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table A11</th>
<th>ERA rating scale 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>Descriptor</td>
</tr>
<tr>
<td>5</td>
<td>UoE profile is characterised by evidence of outstanding performance well above world standard presented by the suite of indicators used for evaluation.</td>
</tr>
<tr>
<td>4</td>
<td>UoE profile is characterised by evidence of performance above world standard presented by the suite of indicators used for evaluation.</td>
</tr>
<tr>
<td>3</td>
<td>UoE profile is characterised by evidence of average performance at world standard presented by the suite of indicators used for evaluation.</td>
</tr>
<tr>
<td>2</td>
<td>UoE profile is characterised by evidence of performance below world standard presented by the suite of indicators used for evaluation.</td>
</tr>
<tr>
<td>1</td>
<td>UoE profile is characterised by evidence of performance well below world standard presented by the suite of indicators used for evaluation.</td>
</tr>
<tr>
<td>n/a</td>
<td>Not assessed due to low volume. The number of research outputs does not meet the volume threshold standard for evaluation in ERA.</td>
</tr>
</tbody>
</table>

mission to facilitate understanding of the research element of the creative work. The research statement follows a standard template and includes:

- research background – field, context and research question
- research contribution – innovation and new knowledge
- research significance – evidence of excellence.

Aggregation of the indicators to produce the overall rating for a UoE submission is through discussion within the expert panel. Individual metrics are also aggregated by discipline in the overall report, giving an indication of the varying contributions of the different subject areas. The data are presented graphically in a variety of forms. Some examples of the
For disciplines where citation analysis is not used, the low volume threshold was increased to 30 submitted research outputs. Following the first full implementation in 2010, one of the main changes in the proposed 2012 exercise is the removal of journal rankings, which had been highly controversial (see, for example, Rowbotham, 2011). Journals had been assigned rankings from A* to C, as an assessment of the quality of the journal. This was felt to be inappropriate, and was thought to lead to perverse incentives, with academics being pressured by research managers to publish in A or A* journals. The new system to be used in 2012 will still include a list of journals with a publication profile for each indicating how often it was chosen as the forum

- A selection of esteem indicators were included.
- There were ERA listed conferences for selected disciplines, which included ranked and unranked conferences.
- Institutions were required to apportion FoR for research outputs, research income, applied measures and esteem. For journal articles and listed conferences, apportionment of FoRs was only possible across the codes to which each journal or conference is assigned.
- For peer review, institutions were able to nominate a preferred FoR in which the output should be reviewed.
- Applied measures were included, covering plant breeders’ rights and NHMRC endorsed guidelines.
for publication by academics in the relevant field. However, the committee members will be allowed to use their own discretion, given their knowledge of the nuances of the field, to assess the quality and appropriateness of journals for different publications and in different research areas. This is something of an adjustment from the originally proposed approach, but does not fundamentally alter the nature and principles of the ERA methodology. Some additional refinements to the methodology have also been suggested, following two public consultations after publication of the 2010 results (ARC, 2012):

- improved capability to accommodate interdisciplinary research – in an extension of the arrangement trialled in 2010 for the mathematical sciences, institutions will be permitted to code a journal article with significant content (66 per cent or greater) not represented by the journal’s FoR(s) to the FoR code that best describes the content
- for peer review disciplines, an increase in the low volume threshold to 50 apportioned weighted outputs, bringing it in line with the threshold for citation disciplines (50 apportioned indexed articles)
- a change in the rules for the attribution of patents, plant breeders’ rights and registered designs to allow them to be submitted when they are granted to eligible researchers (and when they are granted to institutions)
- the modification of fractional staff eligibility requirements so that staff employed at 0.4 FTE

Figure A9
Example of aggregated output from ERA 2010: contribution to national landscape (%)
or greater are automatically eligible, while staff below this threshold are eligible where affiliation is shown (through use of a by-line, for instance).

In addition, the ARC is exploring ways to strengthen the peer review process, though no further details have yet been released as to what these may be.

The report produced on the basis of the data collected in the 2010 ERA process included the following sections, each targeted at a different set of potential uses and users:

• **Section 1, ERA national overview**: a national level summary of a range of ERA data, including research outputs, staff, income received and applied measures

• **Section 2, ERA national profiles by FoR code**: a summary for each FoR code evaluated in ERA, which includes ERA rating statistics and key volume and activity information

• **Section 3, ERA national summary of indicators by FoR codes**: a breakdown of ERA data by two- and four-digit FoR codes

• **Section 4, Australian institution report**: ERA ratings for each two- and four-digit FoR code for each eligible institution.

The data so far have partly been used for accountability purposes, to demonstrate primarily to the Australian public that investment in research has led to success, and also for advocacy purposes, to raise the profile of Australian research internationally. It was stated that in the first round of implementation the outcomes would not have funding implications, which suggests that this may change in the future. It has not been announced at this stage whether the outcomes of ERA 2012 will be used in funding decisions, or if so how, but proposals to use these outcomes in the allocation of Sustainable Research Excellence (SRE) funding are under consultation (Innovation, 2011).

**Analysis**

Given the significant change from the status quo that this represents in Australia, this approach has received remarkably little criticism and has generally been employed effectively and rapidly. Therefore it is likely that this approach will remain for some time to come, changes in political environment notwithstanding, and certainly a further deployment in 2012 is already planned. The reasons for this general acceptability is perhaps related to the context, as this was an alternative to the proposed RQF, which would probably have required significantly more work on the part of the participating institutions and academics, and was highly unpopular. Although ERA does place some burden on participants, in the context of what might have been, perhaps this does not seems so onerous. It is also important to note that Australian universities had already been collecting full publication data in the mid-1990s, which made compliance with ERA requirements relatively straightforward, certainly more so than such a process would be in some other countries.

The approach seems to meet most of its stated aims, giving an overview of the quality of research conducted across the spectrum of institutions and disciplines in Australia, with the ranking scale having been tailored so comparisons and benchmarking on an international level are possible. Results are comparable across time and between institutions and subject areas within them. However, questions inevitably remain around the choice of indicators. With all such approaches the indicators chosen will determine the results obtained, and particularly in the ‘application’ category it is not clear that the indicators included span the range of outcomes that ARC and the Australian government would hope to see from research; rather they reflect those for which it is possible to collect and verify data. Where indicators do not match desired outcomes, there is the potential for perverse incentives to be created, as illustrated by the issues surrounding the use of journal rankings. It is likely that from 2010 to 2012 the process will continue to be developed, and it will be refined in the future.

One possible criticism of the measure of the extent of equality and diversity in research is the use of a low volume threshold, which may disadvantage new areas of research, or institutions that are new to research and may only have a low output in many areas. However, broadly the way in which assessment takes place is at a truly departmental level, with publications submitted as an entire unit and assessed in that way. By not focusing on individual
Much of the criticism of the approach has surrounded the use of journal rankings, which has been modified for the 2012 deployment as described above. However, there was also speculation that the arts and social sciences might receive lower overall rankings comparatively, as they are more dependent on peer review than citations, which is known to be conservative in the ratings it produces. It was thought this may lead to fewer 5s and 4s being awarded, which may affect some HEIs more than others depending on their disciplinary portfolio (see, for example, Trounson, 2011). There has been some criticism around this since the results of the ERA 2010 process were released, but it is not completely clear to what extent this disadvantage is truly present (Wells, 2011). It seems that the social sciences might have been affected by a systematically lower ranking due to peer review (Butler, 2011).

### SWOT analysis

#### Figure A10

A SWOT analysis for ERA

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable to research community in Australia</td>
<td>Indicator driven</td>
</tr>
<tr>
<td>Burden to participants moderate</td>
<td>Still moderated through peer review, reducing objectivity</td>
</tr>
<tr>
<td>Indicator driven</td>
<td>Not comprehensive – academic focus</td>
</tr>
<tr>
<td>Produces a single performance indicator, which can be used for ranking</td>
<td>Summative</td>
</tr>
<tr>
<td>Multi-disciplinary</td>
<td>Burden relative to return is high (not yet used for funding allocation)</td>
</tr>
<tr>
<td></td>
<td>Requires some central expertise (bibliometric expertise on panel)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential to add new indicators</td>
<td>No funding implications</td>
</tr>
<tr>
<td></td>
<td>Politics informed its development</td>
</tr>
<tr>
<td></td>
<td>Government and public appetite to include impact is limited in Australia (limits potential for development)</td>
</tr>
</tbody>
</table>
The National Institute of Health Research Dashboard

Brief description

The Dashboard framework is a conceptual approach for defining key performance indicators that can be effectively linked with longer-term research impacts. The Dashboard provides an empirical link between upstream activities of the health research system and its longer-term strategic objectives. It was developed by RAND Europe for the UK National Institute for Health Research (NIHR) to allow it to meet its reporting requirements and build its monitoring and evaluation capability. The framework is specific to a health research system and was designed to develop tailor-made performance indicators for the NIHR. However, the conceptual methodology is applicable to other national health research systems and has the flexibility of providing context-specific performance indicators.

Origin and rationale

The Dashboard is a conceptual framework that combines aspects of two performance frameworks: the logical frame approach (LFA) and the balanced scorecard (BSC). From the LFA it adopts a logic model to track the longitudinal relationship between investment, activities and outcomes over longer time periods. The performance elements that are borrowed from the BSC are financial performance, internal processes and interactions with external parties. These elements are intended to capture the non-linear relationships of the research process. The key requirements of the framework were to address the long lead time between investment activity and returns on investment that are inherent in health research; the non-linear relationships of the research process; and the effect of this non-linearity on attributing outcomes to research.

The Dashboard methodology aims to develop a small but balanced set of indicators to support strategic decisionmaking by providing regular and structured performance measures. It provides a mechanism for prospective monitoring of the performance of the health research system. The performance indicators focus attention on activities of most importance to an organisation and its stakeholders. The Dashboard is aimed at giving a balanced view of organisational performance while at the same time minimising the burden of data collection for the organisation.

The UK government health research strategy, Best Research for Best Health, is delivered through the NIHR, which was set up in 2006 under the Department of Health. In adopting a central role of maximising the benefits of health research within the English National Health Service (NHS), the NIHR is required to fulfil the UK government’s commitments for monitoring and evaluating its activities and performance to ensure efficient and effective public sector delivery. The Dashboard was developed to help the NIHR fulfil this requirement.

13 The Dashboard framework conceives the NIHR as a national system of health research. It is the only means of implementing the national health research strategy, and it covers the following five components: research priority setting, financing research, research governance, research capacity development and conducting research activities.

14 The ‘learning and growth’ element was not incorporated into the framework because the NIHR already had in place adequate mechanisms for capturing this domain. See El Turabi et al. (2011).
NIHR produced initial implementation plans that set out the specific programmes of activity that would be undertaken to deliver the national research strategy. Indicators were developed for each of these initial implementation plans across the range of logic model stages (inputs, processes, outputs and impacts) and across the breadth of balanced scorecard domains (internal, external and financial domains), as discussed in the next section.

The aim of the dashboard was to produce a quarterly performance report that could be used to support day-to-day management of the NIHR system, with metrics relating to longer-term impacts being reported in a separate NIHR outputs framework.

It also ensures compliance with the pre-existing best practice framework FABRIC for the performance information system developed by the UK government (HM Treasury et al., 2001).

**Scope**

The scope of the NIHR Dashboard work was to develop high-level (system-wide) indicators that covered the entirety of NIHR’s functions as defined by a national health research strategy (Best Research for Best Health), as shown in Table A12.

NIHR was established with the aim of creating an applied health research system embedded in the English NHS. It undertakes all activities of a health research system (priority setting, financing, capacity development, conducting and governing research activity), and as a result the scope of the NIHR Dashboard includes the development of indicators for all aspects of a health research system.

<table>
<thead>
<tr>
<th>Table A12</th>
<th>NIHR implementation plans</th>
</tr>
</thead>
</table>
| **The National Institute for Health Research** | Implementation plan 1.1: NIHR  
Implementation plan 2.1: Funding transition                                              |
| **NIHR faculty**                       | Implementation plan 3.1: NIHR faculty                                                      |
| **Research systems and governance**    | Implementation plan 4.1: Bureaucracy busting: governance, advice and ethics systems  
Implementation plan 4.2: Bureaucracy busting: research information systems                |
| **NHS research infrastructure**        | Implementation plan 5.1: Clinical Research Network for England  
Implementation plan 5.2: Clinical research facilities for experimental medicine  
Implementation plan 5.3: Technology platforms  
Implementation plan 5.4: NIHR School for Primary Care Research                            |
| **NIHR projects, programmes, units and centres** | Implementation plan 6.1: Overview of NIHR research projects, programmes, units and centres  
Implementation plan 6.2: Research for Patient Benefit (RfPB) and Research for Innovation, Speculation and Creativity (RISC) project schemes  
Implementation plan 6.3: Existing R&D programmes  
Implementation plan 6.4: Invention for Innovation Programme  
Implementation plan 6.5: Programme grants for applied research  
Implementation plan 6.6: Research units  
Implementation plan 6.7: Research centres |

---

---

NIHR updates these implementation plans (referred to as briefing documents) every six months.
**Measurement**

The Dashboard framework collects information at programme level on inputs, process, outputs and outcomes (logic model) for each of the three elements – financial, internal and external processes – of a balanced scorecard. The information provided refers to aims, deliverables, metrics and reporting frequency. Figure A11 shows a template for data collection for the financial element. Similar templates are used for the two other BSC elements. The Dashboard framework is applied across the entire NIHR at the programme level each quarter. The information is then collated to provide high-level indicators on health research at the institutional level.

The Dashboard framework combines the four basic elements of a logic model (inputs, process, outputs and outcomes) with the financial, internal process and user satisfaction elements of a balanced scorecard to produce a matrix, which forms the basic structure of the framework, as shown in Figure A12. Each ‘entry’ of the matrix is made up of four ‘dash cards’ or templates that are used to record information at the programme level on aims, deliverables, metrics and reporting frequencies. These programme-specific dashboards provide programme-level performance indicators.

Programme-specific dashboards can be ‘pooled’ to provide a system level dashboard to provide key performance indicators of the health research system, as shown in Figure A13.

The NIHR pilot system-wide dashboard contained 15 indicators distributed along the LFA and BSC elements, as shown in Table A13. They reflected a focus by NIHR on external and early phase effects of research system activity. The 15 indicators were designed to meet the basic properties of the UK government best practice framework FABRIC – the indicators are appropriate, balanced, robust, integrated and cost effective. They are collected quarterly.

**Application to date**

A live Dashboard, slightly modified from the pilot based on feedback, was launched in July 2011 for implementation at NIHR-wide level. The four NIHR coordinating centres will provide data at the programme level of the health research activities that are funded by the Department of Health. Data are

---

**Figure A11**

Template for collecting financial data for the NIHR Dashboard framework

<table>
<thead>
<tr>
<th></th>
<th>Input</th>
<th>Process</th>
<th>Output</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Financial</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aim</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deliverable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes/ actions needed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
is envisaged that ad-hoc analysis will be performed by the NIHR head of business intelligence to ensure proper use of the framework.

**Analysis**

The Dashboard framework has only been recently implemented and the first results were available in mid-2012. An
assessing the effectiveness of this methodology is therefore premature. Aspects such as focus and balance of proposed indicators can only be confirmed through the application of a framework over a period of time. However, one of the main attributes of the conceptual framework is that it is designed to provide a pragmatic approach to defining tailor-made indicators for assessing performance. By focusing on key activities of an organisation or institution, it has the potential to provide targeted and balanced indicators to inform performance relating to strategic objectives. This suggests that the clarity of the strategic objectives plays an important role in developing performance indicators. Equally important, even in cases where strategic objectives are clear, is that the development of adequate performance indicators requires appropriate detailed understanding of the conceptual framework and the underlying complexities of research processes and systems.

**Wider applicability**

The methodology could be applicable to other national health research funders, depending on the wider context. Although tailor-made for the NIHR the broad specifications for design are inherent across
be adapted to research systems in other sectors, using performance indicators appropriate in those sectors. The experiences of developing the NIHR Dashboard suggest that implementation of the dashboard in other contexts would require the following:

- Clear articulation of system-wide strategy and implementation plans is an important pre-condition, but applying the framework in contexts where no such articulation of strategic plans exists is possible. In such circumstances it is important that any performance management framework is developed in close association with the evolving strategy, so the framework can inform strategy development and vice-versa.

### Table A13
Fifteen indicators of the NIHR pilot Dashboard

<table>
<thead>
<tr>
<th>Strategic goal</th>
<th>Indicators</th>
</tr>
</thead>
</table>
| **Goal 1:** Establish the NHS as an internationally recognised centre of research excellence | Number of applications to NIHR faculty and research programmes  
Number of industry funded trials taking place through NIHR networks and experimental medicine research facilities  
Total value of industry funded trials taking place in the NHS  
% of NIHR portfolio trials achieving greater than 95% of recruitment plan  
Number and favourability of news stories relating to NIHR and NHS research in national and international media each month |
| **Goal 2:** Attract, develop and retain the best research professionals to conduct people-based research | Number of faculty members at each level of NIHR faculty (investigators, trainees and associates)  
% of personal award scheme (research fellowship) applicants coming from university departments rated as excellent in the 2008 UK RAE |
| **Goal 3:** Commission research focused on improving health and care           | % of planned research expenditure disbursed  
% of applications for research grants deemed fundable that were funded  
Number of new publications that attribute NIHR funding deposited on UK PubMed Central each month and % of these in journals targeted at practitioners |
| **Goal 4:** Strengthen and streamline systems for research management and governance | Average time from commission to commencement of NIHR portfolio studies  
Number of research passports active  
Pages accessed through the NIHR portal and website |
| **Goal 5:** Act as sound custodians of public money for public good            | % of total cost of programme funding spent on research administration  
Number of risk register issues escalated for monitoring by NIHR Senior Management Team |
• Good communication of the rationale for applying such a novel framework in the selected context is necessary, so health research system stakeholders are ready to support the development of performance indicators.
• There must be adequate training and ongoing support to managers tasked with using the integrated dashcards to develop performance indicators. At pilot stage, managers were encouraged to provide feedback for input in developing the performance indicators.
• A performance information system (IT and human resources) must be identified to manage the collection and analysis of performance information.

**SWOT analysis**

Figure A14  
A SWOT analysis for the NIHR Dashboard

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned with institutional goals</td>
<td>High central burden</td>
</tr>
<tr>
<td>Bespoke</td>
<td>Bespoke</td>
</tr>
<tr>
<td>Formative</td>
<td>Reliant on information management systems</td>
</tr>
<tr>
<td>Can be used for monitoring (frequent assessments)</td>
<td>High up from burden</td>
</tr>
<tr>
<td>Wide applicability</td>
<td>High level of central expertise required</td>
</tr>
<tr>
<td>Strong theoretical basis</td>
<td>Not comprehensive if incorrectly used – it only monitors the indicators you select</td>
</tr>
<tr>
<td>Comparable</td>
<td>Continuous burden (not episodic)</td>
</tr>
<tr>
<td>Focused and selective set of indicators</td>
<td>Not multi-disciplinary</td>
</tr>
<tr>
<td>Indicator set is balanced</td>
<td></td>
</tr>
<tr>
<td>Continuous burden (not episodic)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexibility may allow use across multiple institutions</td>
<td>Scalability across multiple institutions not demonstrated</td>
</tr>
<tr>
<td>Useful at many levels</td>
<td>New and not fully implemented</td>
</tr>
</tbody>
</table>
The Research Excellence Framework

**Brief description**

The Research Excellence Framework (REF) is a research assessment tool which is used to assess the quality of research in UK HEIs. This assessment forms the basis for core quality-related research funding distribution by the Higher Education Funding Council for England (HEFCE), the Scottish Funding Council (SFC), the Higher Education Funding Council for Wales (HEFCW) and the Department for Employment and Learning, Northern Ireland (DELNI). The stated aims of the REF are to:

- inform the selective allocation of research funding to HEIs on the basis of excellence
- provide benchmarking information
- provide accountability for public investment in research and demonstrate its benefits (HEFCE et al. 2009).

Assessment is conducted at the level of subject-specific sub-panels, which roughly correspond to academic disciplines, although funding is allocated by institution. The assessment is made on the basis of three key elements:

- quality of research
- impact of research
- vitality of the research environment.

Assessment is primarily by peer review through ‘expert panels’ supplemented by some quantitative indicators, including bibliometrics in some cases.

**Origin and rationale**

As stated above, the aims of the REF are to inform funding allocation, benchmark institutions and provide accountability for the public money spent on research. However, this does not necessarily give an accurate reflection of the thinking behind this approach and the envisioned benefits when the REF was first proposed. To understand the origins and rationale of the REF, it is important to consider its predecessor, the RAE.

**The RAE**

The primary purpose of the RAE was to produce quality profiles for each submission of research activity made by institutions. The four higher education funding bodies (HEFCE, HEFCW, DELNI and SFC) used the quality profiles to determine their funding allocation for research to the institutions which they fund. Any HEI in the UK that was eligible to receive research funding from one of these bodies was eligible to participate. The first RAE was undertaken in 1986. For the first time it introduced an explicit and formalised assessment process of the quality of research. Further exercises were held in 1989, 1992, 1996, 2001 and 2008. Assessment was exclusively on research quality, with ratings awarded by peer review by a panel of subject-specific experts on this scale (which applies to 2008; in previous years, the ratings scale differed) (RAE, 2008):

- **four star**: quality that is world leading in its originality, significance and rigour
- **three star**: quality that is internationally excellent in its originality, significance and rigour but nonetheless falls short of the highest standards of excellence
- **two star**: quality that is recognised internationally in its originality, significance and rigour
• **one star**: quality that is recognised nationally in its originality, significance and rigour
• **unclassified**: quality that falls below the standard of nationally recognised work, or work that does not meet the published definition of research for the purposes of this assessment.

**A move towards the REF framework**

Initially, the intended primary objective of the new system which would become the REF was to reduce the level of burden and cost of the assessment process, which was perceived to be too high (HM Treasury et al., 2006). The intention was to achieve this through the use of quantitative (primarily bibliometric) indicators to largely replace or at least minimise the necessity of peer review for assessing research quality. A bibliometrics-based process was expected to be much cheaper, and less burdensome (HM Treasury et al., 2006). The stated aims of the new evaluation framework given in a letter circulated to HEIs in 2007 were to:

• produce robust UK-wide indicators of research excellence for all disciplines, which can be used to benchmark quality against international standards and to drive the Council’s funding for research
• provide a basis for distributing funding primarily by reference to research excellence, and to fund excellent research in all its forms wherever it is found
• reduce significantly the administrative burden on institutions in comparison to the RAE
• avoid creating any undesirable behavioural incentives
• promote equality and diversity
• provide a stable framework for our continuing support of a world leading research base within HE (HEFCE, 2007).

However, following fairly widespread criticisms from within the academic community (REF, 2008) and a bibliometrics pilot exercise (HEFCE, 2009a), there has been significant backtracking on this proposal. This applies particularly to the intended use of bibliometrics. For example, there were significant concerns among the academic community about the potential implications of assessing the science-related and non-science-related subjects differently, as proposed under the bibliometric approach.

As a result, the current REF framework proposals are in essence an extended RAE, with an additional significant assessment component accounting for 20 per cent of the overall evaluation covering the wider impacts of research outside academia, such as those on society and the economy.

Whether the REF reduces burden, or even aims to do so, is questionable, although this was one of the key rationales for making changes to its predecessor, the RAE. Some measures, such as reductions in the number of separate units of assessment, may reduce burden, but the inclusion of impact case studies will increase burden. However, the REF does reflect some changes in attitudes towards research assessment by comparison with the RAE. Perhaps, given increasing financial pressures, the aim of the REF could be considered to include the need to make the case for spending of public money on research. By including wider impact elements in the assessment, academics are given the opportunity to demonstrate why their research matters. Higher education funding councils are given a stronger case to present when defending their budgets. This certainly reflects wider moves to accountability in research assessment, both in the UK and more widely, as the need to show the societal and economic return to taxpayers on their investment in research grows. This is attempted through a broader means of assessment, incorporating not just academic research quality, but factors covering non-academic impacts and the wider research environment.

Another important influence on the development of the REF was the RQF, a research evaluation approach developed in Australia but never employed, largely for political reasons. The methodology and proposed aims are fundamentally similar, and the thinking behind the RQF was influential in the development of the REF approach as it is currently envisaged.

**Scope**

The REF will provide assessment, leading to funding allocation, for all 165 HEIs in the UK from 2014. A set of 36 sub-panels have been established, spanning all subject areas. HEIs will make submis-
sions to as many of these as they deem relevant to their research interests, and assessment will be made at the sub-panel level. However, funding allocation will be made by HEI, taking into account the HEIs scores from each sub-panel. The number of sub-panels is reduced significantly compared with the RAE (which had 67 sub-panels in 2008), the principle being that this will reduce burden and spread it more evenly, as submission levels varied significantly between different sub-panels in the RAE. It is thought that this may also better facilitate the assessment of multi- and cross-disciplinary research.

Assessment is carried out at the sub-panel level, but this is under the guidance of four overarching main panels (labelled A to D, broadly corresponding to the biological sciences, physical sciences, social sciences, and arts and humanities, respectively), which help to ensure consistency between sub-panels and provide guidance and arbitration where required. Again, the number of main panels is significantly lower than for the RAE, and it is intended that this may produce better consistency of assessment. HEIs may make multiple submissions to the same sub-panel where necessary, though this is discouraged.

The research is assessed on the basis of three elements: quality of research outputs, impact of research, and vitality of the research environment:

- Quality is the dominant element of the assessment, and the criteria to be used are ‘originality, rigour and significance’ in assessing the academic excellence of the research outputs, such as journal publications or other academic texts.
- The impact element covers all non-academic impacts of the research, spanning everything from societal benefits, such as improvements in healthcare, to economic gains, such as the creation of new products. The criteria here are ‘reach and significance’.
- The vitality element of the assessment aims to reward research environments that support high-quality research and dissemination or application. This covers teaching within the HEI and academic dissemination, and the criteria are ‘vitality and sustainability’.

Currently the REF is at the implementation stage, having undergone significant modifications through consultation and the early pilots, particularly in the use of bibliometric data. The first full implementation of the REF is scheduled for 2014, and preparations are well under way centrally and at HEIs.

Measurement

There are three elements which are assessed under the REF. Here we outline the data collected and assessment process for each of them, the format of the expert panels, and the way in which the three sets of data will be aggregated.

Quality of research outputs

This is the dominant element of the assessment. In current proposals it will account for 65 per cent of the final assessment, although this is likely to be reduced to 60 per cent in future iterations when the other elements of the assessment, in particular the impact element, are more developed.

HEIs select which staff and which of their outputs to submit for assessment, with each member of staff selecting their four ‘best’ publications over the period of evaluation (2008–2013). This can be supplemented by some degree of bibliometric data, the extent of which is to be determined by each sub-panel. The overarching guidance is that bibliometrics should only be used for subject areas where there are known to be ‘robust’ indicators. In practice this means, broadly, that bibliometric indicators will only be applied in STEM subjects, and treated as secondary to the peer review of submissions in all cases. For example, in Panel A, which broadly covers the biological sciences, the current position is that all sub-panels will use citation data as part of their assessment (HEFCE et al., 2011c), while in Panel D, broadly covering the arts and humanities, the converse is true, with no sub-panel using citation data (HEFCE et al., 2011f). The picture for panels B and C (roughly corresponding to the physical sciences and the social sciences respectively) falls somewhere between the two, with some sub-panels choosing not to use the citation data and some sub-panels using the data, to differing extents (HEFCE et al., 2011d, 2011e). This is a significant change from the origi-
nal proposals, in which bibliometrics were intended to form the core of assessment in this area, as mentioned previously, and outlined in more detail below. Bibliometrics are to be provided centrally by HEFCE where used. All but one sub-panel will use only this bibliometric data, if any. One sub-panel may supplement this with data from Google Scholar.

The criteria for the assessment of these academic outputs will be ‘originality, rigour and significance’, and assessment will be via peer review by a subject-specific expert panel. The specifics of this expert panel are outlined in more detail below. Ratings will be made on a scale from 1* to 4*, and are intended to be internationally benchmarked, so 4* corresponds to world leading research. The ratings are intended to be broadly comparable to the previous RAE ratings system in 2008, although the descriptor for the highest 4* rating has been revised slightly and is intended to be more discriminating of the very best research. This is the proposed ratings scale:

- **four star**: ‘quality that is world leading in terms of originality, significance and rigour’
- **three star**: ‘quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence’
- **two star**: ‘quality that is recognised internationally in terms of originality, significance and rigour’
- **one star**: ‘quality that is recognised nationally in terms of originality, significance and rigour’
- **unclassified**: ‘quality that falls below the standard of nationally recognised work, or work which does not meet the published definition of research for the purposes of this assessment’ (HEFCE, 2009b).

**Impact of research**

The impact of research was initially intended to account for 25 per cent of the total assessment, but this has been scaled back, at least for this first implementation, to 20 per cent. This reflects the level of uncertainty over how this element will be assessed and, to some extent, concern of academics in particular research fields. However, it is likely that the proportion will be increased to 25 per cent in future.

For the purposes of the REF, impact is defined as an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life beyond academia (HEFCE et al., 2011a). For example, in the health sciences, research may have developed new medical treatments, leading to impacts on health, and possibly also economic impacts through improved efficiency, or through development of business in production of drugs or devices. In the humanities, research might have helped to preserve cultural heritage, such as a language or artefact. In the social sciences, research might have contributed to public debate, either through media engagement, or contributed to debate at a political level, through providing evidence to government, thereby changing policy, or shaping and contributing to discussion and debate. Further examples of eligible impacts are provided in the specific guidance published by the four main panels (HEFCE et al., 2011c, 2011d, 2011e, 2011f, 2012).

The aim of this element of the assessment is to reward not just the research that is of the highest quality, but also that which is of great utility. This reflects a growing need for public accountability in research funding.

Impact is to be assessed using case studies of specific ‘research-driven impacts’ and overarching ‘impact statements’, which describe the breadth and range of interactions with research users and the effects or outcomes of these interactions. The criteria for assessment of case studies are ‘reach and significance’, where reach is defined as how widely the impacts have been felt, and significance is defined as how transformative the impacts have been. Case studies will describe specific examples of impact resulting from research, while impact statements will also have a units overarching approach to enabling impact from its research. HEIs will submit one impact statement per sub-panel submission, plus case studies. The number of these depends on the number of research staff in the submission, as shown in Table A14.

Impact must have occurred during the REF assessment period (2008–2013), but can be based on research undertaken earlier (1993–2013), reflecting the time lag from research to impact. A wide
range of impacts are eligible. Examples of appropriate impacts are provided at the main panel level in the guidance (HEFCE et al., 2011c, 2011d, 2011e, 2011f), but they must be non-academic, so changes in teaching within the HEI, for example, would not be eligible. HEIs must also include evidence that the impacts are a direct result of the research as they claim, describing the relationship through the case study narrative and providing hard evidence of impact where possible. Requirements for quantitative rather than qualitative evidence differ between the main panels. However, evidence of fact (figures showing increased turnover, user numbers and demographics for a cultural event) are preferred to evidence of opinion (such as testimony of personal opinion from specific users or collaborators). Impacts do not travel with academics: institutions may claim impact from work conducted by staff that have left, provided the institution is still active in that area of research, but conversely institutions may not claim impacts from research conducted by staff before starting work at that institution.

The outcomes of the impact assessment are used to produce an impact sub-profile for each submission, which reflects the proportion of each submission that reaches a particular standard on a ratings scale from 1* to 4*, with an additional rating of ‘unclassified’ available. Each of these ratings carries a generic definition for the expected level of reach and significance. The criteria for assessing impacts are ‘reach and significance’. It is intended that in assessing case studies, sub-panels will form an overall view about the case study’s ‘reach and significance’ taken as a whole, rather than assess reach and significance separately. Similarly, in assessing the impact template the panel will consider the extent to which the unit’s approach described in the template is conducive to achieving impacts of ‘reach and significance’ (HEFCE et al., 2011b). These are the definitions of the rating scales:

- **four star**: ‘outstanding impacts in terms of their reach and significance’
- **three star**: ‘very considerable impacts in terms of their reach and significance’
- **two star**: ‘considerable impacts in terms of their reach and significance’
- **one star**: ‘recognised but modest impacts in terms of their reach and significance’
- **unclassified**: ‘the impact is of little or no reach and significance; or the impact was not eligible; or the impact was not underpinned by excellent research produced by the submitted unit’ (HEFCE et al., 2011a).

What this means in practice is outlined in more detail in the panel-specific guidance (HEFCE et al., 2011c, 2011d, 2011e, 2011f).

**Vitality of the research environment**

The aim of this element of the assessment is to reward HEIs which have a research environment that supports high-quality research and dissemination or application. It is weighted at 15 per cent of the total assessment. Submissions are made via a standardised template, aiding comparison. The template collects data in the following sections:

- overview
- research strategy
- people, including:
  - staffing strategy and staff development
  - research students
- income, infrastructure and facilities
- collaboration and contribution to the discipline.

It is intended that submissions will be largely qualitative in nature, supported by some key quantitative indicators, covering the number of research doctoral degrees awarded, research income (broken down by source) and research income in kind. These quan-

<table>
<thead>
<tr>
<th>Number of staff submitted (FTE)</th>
<th>Number of case studies required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 14.99</td>
<td>2</td>
</tr>
<tr>
<td>15–24.99</td>
<td>3</td>
</tr>
<tr>
<td>25–34.99</td>
<td>4</td>
</tr>
<tr>
<td>35–44.99</td>
<td>5</td>
</tr>
<tr>
<td>45 or more</td>
<td>6, plus 1 further case study per additional 10 FTE</td>
</tr>
</tbody>
</table>

SOURCE: HEFCE et al. (2011a)
Quantitative data will be collected in parallel with data which are already collected for other purposes and hence should not add extra burden.

The assessment criteria for the research environment are ‘vitality and sustainability’, covering not only the ‘vitality and sustainability’ of the submitted unit, but also its contribution to the ‘vitality and sustainability’ of the wider research base. These are the generic statements for the five different levels:

- **four star**: ‘an environment that is conducive to producing research of world leading quality, in terms of its vitality and sustainability’
- **three star**: ‘an environment that is conducive to producing research of internationally excellent quality, in terms of its vitality and sustainability’
- **two star**: ‘an environment that is conducive to producing research of internationally recognised quality, in terms of its vitality and sustainability’
- **one star**: ‘an environment that is conducive to producing research of nationally recognised quality, in terms of its vitality and sustainability’
- **unclassified**: ‘an environment that is not conducive to producing research of nationally recognised quality’ (HEFCE et al., 2011a).

It is interesting to note that this element of the assessment has been significantly less controversial than the impact section and the use of bibliometrics, both of which have been the subject of a pilot and subject to differing levels of revision as a consequence. The academic community seems comfortable with this element of assessment and what it may entail, as reflected in the consultation feedback. This may be because these data are collected at a unit, rather than individual, academic level – as individual academics do not have to engage with this step, and no demands are made on their time, they are less likely to make complaints. In consultation, there was ‘widespread agreement that the research environment should be a significant factor in the assessment’. Responses stressed the importance of including future strategies, data on research income and students, training and development of staff, evidence of wider activity to support the UK research base, and engagement with research ‘users’ and the wider public in the assessment (although there was a desire for increased clarity on where this fits with inclusion of this information in the impact statement) (HEFCE, 2010).

**Expert panels**

Assessment is carried out by subject-specific sub-panels under the guidance of four overarching main panels, with multiple submissions to the same sub-panel by one institution possible (but discouraged). There are more members in each sub-panel than for the RAE and sub-panels are able to recruit additional members if required to cover the breadth of subject knowledge required. As well as conducting the assessment, the panels are also responsible for developing the assessment criteria and the methods that the working panels use. In particular, the main panel provides leadership and guidance to a group of sub-panels, while the sub-panels conduct the assessment and define specific methods appropriate to their subject area. Main panels are also responsible for coordinating and ensuring consistency between assessments from the different sub-panels.

Main panels are composed of a chair, who is recruited through an open process, the chairs of each corresponding sub-panel, and additional members with international experience, or expertise in the wider use and application of research in that area.

Sub-panels are expected to have between ten and 30 members, depending on the breadth of the sub-panel’s remit. Members are drawn from highly respected academics, and appropriate representatives from private, public or third sectors with expertise in commissioning, applying or making use of research. The purpose of this second group of members is particularly to provide input to the ‘impact’ component of the assessment. Although not all users are involved from inception, at least one research user is appointed to each sub-panel early enough to be involved in the development of assessment criteria.

**Final outcomes**

The final outcome for each submission consists of a sub-profile for each of the three elements outlining the proportion of submitted work meeting each level in the five-point scale suggested, along with an overall excellence profile, which combines the three sub-profiles. In 2014 the sub-profiles will be combined using the following weightings:
• output: 65 per cent
• impact: 20 per cent
• environment: 15 per cent (REF, 2011).

It is expected that the weighting for the impact element will rise to 25 per cent after 2014, reducing the output component to 60 per cent in the future, though this will depend on the success of the impact assessment in this first deployment. Details of how the sub-profiles will be combined are provided in Annex B of the guidance on submissions (HEFCE et al., 2011a).

The REF replaces the RAE, which was last run in 2008 and was a regular process operating on a six-yearly cycle. It is reasonable to assume that the REF will also take place with similar regularity, subject to its performance on this first outing; however, this is yet to be confirmed by HEFCE or others. The first REF process is underway, with the assessment scheduled to take place in 2014. Further information about the timelines of this assessment is outlined below.

Considering the level of replicability of this process brings into consideration arguments about the pros and cons of peer review more widely as an assessment process. There is a substantial literature around this, with differing opinions on the level of reproducibility between assessors (Cole et al., 1981; Hodgson, 1997; and Jayasinghe, Marsh and Bond, 2003) (best estimates suggest reproducibility is around 75 per cent for panels, but significantly lower for individual reviewers), and a range of other criticisms raised, such as arguments that peer review processes are conservative, and can disincentivise new thinking (Roy, 1985; and Braben, 2004). However, these arguments, and indeed the arguments in favour of peer review, have been discussed extensively elsewhere (for example by Ismail, Farrands and Wooding, 2009). Ultimately, peer review remains the gold standard for the assessment of research.

In the specific context of the REF and the RAE, there has been some query made regarding bias. Butler and MacAllister (2009) found that departments that had a member on the assessment panel in the RAE performed better, but this may also reflect the fact that eminent researchers are asked to join the assessment panels, and they are more likely to be a member of a high performing department. It can be assumed that this method is as appropriate and/or replicable as most assessment methods used widely in academia.

Applicability to date

To date, full application of the REF evaluation methodology has not taken place. Indeed, some of the methodology, including the finalised guidance on impact assessment, is still under consultation, although most of the key elements of the proposed approach have been finalised. The impact element of the assessment has undergone pilot testing across five sub-panels (clinical medicine, physics, earth systems and environmental sciences, social work and social policy, and English language and literature) with 29 HEIs participating. The report on these pilots from the panel chairs (HEFCE et al., 2010) and an external assessor looking at the feedback from participating HEIs (Technopolis Group, 2010) concludes that this is a feasible approach, though the experience of the pilot led to significant changes to the original proposals. These include changes to the template for case studies, improvements to the guidance provided to HEIs, limiting the impact element of the assessment to 20 per cent of the full assessment in this first deployment, and improvements in the quality of evidence provided by HEIs. These suggestions have largely been adopted by HEFCE.

Inasmuch as the quality part of the assessment is very similar to the approach used in the RAE, this part of the assessment is well established. What is more novel is the (limited) use of bibliometrics as an input to this process. This method was piloted, and it was found that ‘Bibliometrics are not sufficiently robust at this stage to be used formulaically or to replace expert review in the REF. However there is considerable scope for citation information to be used to inform expert review’ (HEFCE, 2009a). Therefore, current plans are for bibliometric indicators to be made available to sub-panels to use in their assessment, where desired. At present, it appears that this will most likely be the case primarily in science, technology, engineering and mathematics, and has been determined by the sub-panels themselves, on the basis of the relevance of bibliometric indicators in that discipline.
The ‘vitality of the research environment’ component of the assessment has not yet been piloted or deployed. This is perhaps partly because similar data were collected under the RAE.

The timeline for the first deployment of the REF as determined by HEFCE is outlined in Table A15. It is interesting that the assessment period from the closing date of submissions to publication of final outcomes is well over a year. It may be that this time period can be reduced once the evaluation framework is better established.

There are a number of other modifications which are possible in subsequent REF evaluations. Most notably, the impact element of the assessment is likely to account for 25 per cent of the assessment in subsequent deployments, subject to it proving successful in this first evaluation.

Application is limited to the UK, and this is likely to remain the case, although similar approaches may be adopted elsewhere depending on the outcomes of this first round of the REF. Similar proposals had been made in Australia: the Australian RQF to a

Table A15
Timeline for deployment of the REF

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2010</td>
<td>Publication of ‘initial decisions’ on the REF</td>
</tr>
<tr>
<td>July 2010</td>
<td>Publication of ‘Units of Assessment and Recruitment of Expert Panels’ (REF, 2010b)</td>
</tr>
<tr>
<td>November 2010</td>
<td>Publication of reports on the REF impact pilot exercise</td>
</tr>
<tr>
<td>February 2011</td>
<td>Panel membership announced</td>
</tr>
<tr>
<td>July 2011</td>
<td>Publication of equalities briefing for panels</td>
</tr>
<tr>
<td>July 2011</td>
<td>Publication of guidance on submissions</td>
</tr>
<tr>
<td>July 2011</td>
<td>Publication of draft panel criteria and working methods for consultation</td>
</tr>
<tr>
<td>Summer 2011</td>
<td>Consultation on panel criteria and working methods, including workshops for research users to input into the impact criteria</td>
</tr>
<tr>
<td>October 2011</td>
<td>Close of consultation on panel criteria and working methods</td>
</tr>
<tr>
<td>January 2012</td>
<td>Publication of panel criteria and working methods</td>
</tr>
<tr>
<td>Autumn 2012</td>
<td>Pilot of the submissions system</td>
</tr>
<tr>
<td>January 2013</td>
<td>Invitation to HEIs to make submissions; release of submission system and manuals</td>
</tr>
<tr>
<td>Spring 2013</td>
<td>Survey of HEIs’ submission intentions</td>
</tr>
<tr>
<td>31 July 2013</td>
<td>End of assessment period (for research impacts, the research environment and data about research income and postgraduate students)</td>
</tr>
<tr>
<td>Autumn 2013</td>
<td>Appointment of additional assessors to panels (research users and practicing researchers)</td>
</tr>
<tr>
<td>31 October 2013</td>
<td>Census date for staff eligible for selection</td>
</tr>
<tr>
<td>29 November 2013</td>
<td>Closing date for submissions</td>
</tr>
<tr>
<td>31 December 2013</td>
<td>End of publication period (cut-off point for publication of research outputs, and for outputs underpinning impact case studies)</td>
</tr>
<tr>
<td>2014</td>
<td>Panels assess submissions</td>
</tr>
<tr>
<td>December 2014</td>
<td>Publication of outcomes</td>
</tr>
<tr>
<td>Spring 2015</td>
<td>Publication of submissions, panel overview reports and sub-profiles</td>
</tr>
</tbody>
</table>
large extent provided the basis for the approach ultimately developed for the REF (RQF Development Advisory Group, 2006; Grant et al., 2009). The RQF was abandoned in Australia after piloting following a change of government, but this was because of political rather than feasibility concerns, reflecting a different ideology and approach to research evaluation in the incoming government.

It is clear that significant changes were made to the methodology used in the REF, as already discussed. These were the initial proposals, reflecting a government decision to move to a low burden, metrics-based system: ‘The approach for the science-based disciplines (science, engineering, technology and medicine) will be based on quantitative indicators, including bibliometric indicators of research quality and impact, external research income and postgraduate student activity... Research quality in the arts, humanities, social sciences, and mathematics and statistics will be assessed through a light-touch process, based on peer review and informed by statistical indicators in common with the science-based disciplines’ (HEFCE, 2007).

However, following consultation the metrics-based approach has largely been abandoned, and with it the likelihood of reducing burden to any significant extent, but the concept of including wider factors other than academic quality of outputs in the assessment, such as impact and factors relating to the research environment, has been preserved. These changes were made on the basis of the pilots conducted and the wider critical reception in the academic community in the UK, as evidenced in two consultations (DfE, 2006).

Analysis

Assessment of the effectiveness of this approach depends on which set of aims you compare against. When taken in the context of the initial set of aims outlined in 2007, a key tenet of which was to reduce burden, the approach is unlikely to reach its goals. It seems inevitable that the addition of extra assessment components relating to impact and vitality of the research environment will lead to increased cost and burden, however carefully they are structured. Despite restructuring of the panels and possible reductions in the number of submissions required, it seems unlikely that the quality element of the assessment can be significantly less burdensome to assessors and participants than the RAE.

Assessment against the currently stated aims of the process can be more favourable. It seems likely that this measurement approach will produce a comparable rating across HEIs that will allow benchmarking and facilitate the allocation of research funding, and these results will be comparable to subsequent assessments, and broadly comparable to the RAE, at least in their quality element. Furthermore, the introduction of components relating to the wider impact of research are likely to increase levels of accountability for public investment in research. In turn, this may help strengthen the argument for spend on research at a time of economic hardship by providing a portfolio of examples of the benefits research can bring, through the case studies provided during assessment. Therefore, in the broader sense, it seems likely that the approach is fit for purpose, although the first implementation will be the real test of this.

However, it is clear that the approach that is still being finalised is not going to be as radical a departure from the RAE as initially promised. Indeed it can be argued that the REF is really just a modified or extended RAE following the significant compromises that had to be made for political and practical reasons, with the caveat that the introduction of impact, although at low weighting (20 per cent), is still novel and is considered to be a new element of evaluation. Further use of this approach in some form is highly likely, and it is clear that there is no appetite among the UK academic community for a radically changed approach to this process, however burdensome the current system may be. Changes are possible depending on how successful this first deployment proves. Any such changes are likely to be incremental, however, given that this process was supposed to be a step change, but changes ultimately ended up being a much less significant departure from the status quo.

The REF provides specific guidance on equality and diversity to the panels on the quality and environment elements of the assessment. The guidance allows for the number of outputs submitted in the quality element of the assessment to be reduced
from four where circumstances have constrained the ability of a researcher to produce the four outputs. Where these circumstances are ‘clearly defined’ (e.g. for part-time working, early career researchers) there are standard methods defined to work out the appropriate number of outputs required. Where circumstances are more complex (e.g. disability, long-term illness, caring responsibilities), cases are considered in a consistent manner by the Equalities and Diversity Advisory Panel, and recommendations passed to the appropriate Main Panel, which takes a final decision on the reduction in outputs. Submissions with fewer than four outputs to be assessed without penalty and the decision on the appropriate number of outputs where this is reduced should be taken completely separately from the decision on quality. The conditions under which these two different mechanisms for a reduction in the required number of outputs are applicable, and the process by which this reduction will take place, are detailed in the panel criteria and working methods document (HEFCE et al., 2012). One section of the research environment template refers specifically to ‘people’ and in this section institutions are required to provide evidence on how they have promoted equality and diversity. More specific information on what this includes is provided for individual panels.

Outstanding challenges for the REF approach include developing a methodology to measure impact effectively in a way that is widely applicable and acceptable, and how to reduce burden in a system that is so heavily reliant on peer review. The first of these is still being tackled through the analysis of the pilot process, and is likely to be subject to further improvement following assessment of the REF in 2014.

**Wider applicability**

This approach could readily be transferred to similar cross-institutional assessments of performance in some countries, particularly in Europe and/or Australasia, where there is experience of such evaluation systems at the national level. It could be employed by funders supporting research in one specific discipline, as assessment is at the subject area level. However, given the level of complexity of preparation for assessment, use in its current form would require the evaluation to correspond to significant financial allocations. It seems unlikely that this could easily be scaled back without quality of assessment being compromised, as the preparation of accurate and meaningful case studies on impact, or thorough peer review of research quality, is inherently complex and time consuming. Most information gathered in this evaluation is best used for benchmarking (and funding allocation) or for advocacy, but the format in which the evaluation is conducted means the output is not very applicable to internal learning. The exception to this is the research environment component of the assessment, which could provide some internal learning and feed into development of strategy at a number of levels within HEIs. However, the literature around this is not well developed, and it will be interesting to see what this includes when further guidance, including the template for this element of assessment, is released.

Implementation has not been without its challenges, but part of the cost and complexity has resulted from the simultaneous development of the methodology. The lack of clear guidance has led to uncertainty among HEIs and hence increased some of the workload in preparation for the REF as HEIs struggle to adapt to changing goalposts. It is likely that once the rules are clearly laid out, and HEIs have experience in how to produce their REF submissions, this will be easier. It is certainly likely that there will be much transferrable learning at the assessor and HEI level, which would make it easier to apply this type of framework again for another assessor. HEFCE has suggested that transaction costs are likely to be similar to those for the RAE in 2008, which were around £60 million, with £48 million of this being costs to HEIs across the sector (PA Consulting, 2009), but it is hard to see how this will be possible given that the REF is effectively the RAE plus additional assessment elements. Adaptations such as reductions in the number of sub-panels are unlikely to be able to compensate for this additional workload, even if the work associated with a new format for submission and the associated learning are not taken into account. However, some argue that the RAE, and consequently the REF, is not as expensive as often suggested when taken in
context of the amount of funds allocated. The 2008 RAE process may have cost around £60 million to administer, including costs to the participating HEIs, but it led to the distribution of at least £6 billion in quality research funding, so this cost was less than 1 per cent of the total allocation. By comparison, the administration costs of the UK research councils, publicly funded agencies responsible for coordinating and funding research, are 4 per cent of their total budget (RCUK, n.d.), not taking into account the costs to HEIs.

As outlined previously, a range of adaptations had to be made to the additional proposals, partly on the basis of the critical response, focusing particularly on the use of bibliometrics and an overall indicator-focused approach. Following these adaptations, the framework seems overall to have a fairly good level of buy-in from the stakeholder community. However, there has been some additional critical feedback on the more recent version of the proposed REF. Concerns tend to be around the ability of the approach to assess the wide range of different types of research conducted effectively, and in particular whether the move to include impact as an assessment component may threaten basic research. Bekhradnia (2009) suggests that the component of assessment relating to the research environment will favour high performing (and larger) institutions, although his overall assessment of the REF is positive. Some, such as Oppenheim (2008), have criticised the backtracking from the bibliometrics approach, suggesting bibliometrics can be an effective approach and is certainly much less expensive than peer review. However, he does concede that it is a difficult sell to researchers, an issue reflected by others, including Russell (2009), who concludes that peer review is the only method of assessment that academics trust (at least in the UK).

**SWOT analysis**

Figure A15
A SWOT analysis for REF

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burden relative to return is low (determines significant funding allocation)</td>
<td>Cost</td>
</tr>
<tr>
<td>Acceptable to UK academic community as it uses peer review</td>
<td>Total burden on participants and centrally is high</td>
</tr>
<tr>
<td>Comprehensive (includes impact)</td>
<td>Can discriminate against some types of researchers (though efforts have been made to avoid this)</td>
</tr>
<tr>
<td>Multi-method</td>
<td>Can discriminate against some types of institution</td>
</tr>
<tr>
<td>Multi-disciplinary</td>
<td>Summative</td>
</tr>
<tr>
<td>Successfully piloted, and many elements well tested</td>
<td>Scalability not demonstrated</td>
</tr>
<tr>
<td>Produces a single performance indicator which can be used for ranking</td>
<td>Not transparent</td>
</tr>
<tr>
<td></td>
<td>Almost solely reliant on peer review – limits objectivity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential to move towards indicators</td>
<td>Non-participation</td>
</tr>
<tr>
<td>Move towards impact in UK and internationally</td>
<td>Political</td>
</tr>
<tr>
<td>Increased focus on public accountability in UK</td>
<td>Reductions in research funding may limit ability to fund to match the quality demonstrated</td>
</tr>
<tr>
<td></td>
<td>Could result in research concentration</td>
</tr>
</tbody>
</table>

---

16 QR funding from HEFCE alone for the year 2012/13 will be around £1.6 billion. See HEFCE (2012).
Productive Interactions

**Brief description**

Productive Interactions is an approach to research evaluation that is being developed through a European Commission funded project, SIAMPI, by a consortium of researchers (NWO, n.d.). 'SIAMPI' is an acronym for the project, standing for 'Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions between science and society', the name of a study that developed the approach (Spaapen and Van Drooge, 2011).

The use of Productive Interactions as a proxy for research impact is unique to this approach. The premise is that knowledge develops and impact is achieved through a series of interactions between researchers and society. This is modelled as a two-way process, with three types of productive interaction identified (Spaapen and Van Drooge, 2011):

- **Direct personal contacts:** any direct communication, and might be as simple as a conversation but also covers more complex interactions such as research collaborations.
- **Indirect interaction:** any type of interaction that is mediated by a carrier. This could be interaction via a publication of any type, from journal article to clinical guideline, or through the media, an exhibition, film or production, or through artefacts, such as websites, prototypes, demonstrations and designs.
- **Financial interaction:** where there is some kind of economic exchange between stakeholders and researchers. This could take the form of direct research funding, but could also include IP rights, interaction in relation to contracts, or consulting, for example, or interactions through financial contributions in kind, such as facility sharing.

There are five institutions or groups of institution involved in the SIAMPI consortium: the Royal Netherlands Academy of Arts and Sciences (KNAW); the Rathenau Institute (Netherlands); Ingenio, an institute of the Universidad Politénica de Valencia (UPV) and the Consejo Superior de Investigaciones Científicas (CSIC) (Spain); the Maisons de Sciences de l’Homme (France); and the University of Manchester (UK). KNAW is the coordinator for the project and is a stakeholder, but is not involved directly in research. The other institutions are all involved in SIAMPI in a research capacity.

**Glossary**

**ERiC:** Evaluating Research in Context is an approach to evaluating research impact which was developed for use in the Netherlands. The concept of Productive Interactions and the SIAMPI project emerged from the work in developing ERiC.

**FP7:** The European Commission’s seventh framework programme provides research funding across a wide range of areas. The SIAMPI project was developed in response to a particular call through the FP7 Science in Society programme.

**Productive Interactions:** Interactions between researchers and society which are ‘productive’, leading to efforts by the stakeholders to apply research findings and hence change their behaviour in some way. These are interactions that could ultimately lead to impact or benefit to society, and are used as proxies for impact in this approach.
**Origin and rationale**

The SIAMPI project evolved out of the thinking and discussion brought about through the work on an earlier research evaluation framework, Evaluating Research in Context (ERiC), which aimed to provide a method for researchers to understand their impact. ERiC consists of a four-step process (Spaapen, Dijstelbloem and Wamelink, 2007):

- Establish self-image or mission of institute.
- Gather data regarding the impact of the group more widely.
- Identify and contact stakeholders.
- Compare findings to goals and missions, and develop future plans.

In the use of this reflective approach, ERiC and the Productive Interactions draw on the Realist Evaluation approaches (Pawson and Tilley, 1997). An important element of the ERiC approach was that the outcomes of the evaluation should be reflection and learning for the research group, rather than any external or comparative assessment, and this approach persists in the SIAMPI project. However, ERiC did not focus on Productive Interactions, and this concept was developed in one of the later meetings of the ERiC researchers and others. From this existing collaboration and the concept of Productive Interactions, the SIAMPI project developed on the basis of discussions at a workshop in 2007. As well as the focus on Productive Interactions, the other main difference of the SIAMPI project from ERiC is its scope. ERiC was primarily designed to provide a framework for assessing impact as part of the Dutch Standard Evaluation Protocol (SEP), which is the process by which universities are assessed in the Netherlands. For more information on the SEP, see Association of Universities in the Netherlands (n.d.). The Production Interactions approach developed in the SIAMPI project, by contrast, has been developed by researchers from a range of European countries and is designed to be suitable for use in a wide range of contexts. The aim is that it should be an approach which can be applied ex-ante or ex-post, and for a project, programme or group. Many of the Dutch stakeholders involved in the ERiC work found the Productive Interactions concept interesting and useful, so this idea is discussed in some of the later versions of ERiC. However, the idea of using Productive Interactions as a proxy for impact stems from the work of SIAMPI and was not part of the original ERiC work.

The SIAMPI project is supported by the European Commission’s seventh framework programme (FP7) and the team responded to an open call under the Science and Society programme for research into the social impact assessment of research. They received a grant worth €800,000 over two years to support their work in developing the framework, which came to a close in 2011.

The key tenets of the SIAMPI project are two-fold. First, to develop the concept of Productive Interactions: to analyse the process through which impact is realised, rather than just the impact itself, the idea being that this is what researchers and research institutions can control and develop strategically. This also has a secondary advantage, as impact generally takes place long after research is conducted (current estimates in biomedical and health research suggest that this time lag might be in the order of 20 years, and this may differ significantly between fields), whereas interactions can be measured as they happen. Second, and common to ERiC, in the Productive Interactions approach, evaluation should be used for learning and improvement, rather than for judging or ranking. Evaluation is based on aims and criteria, which are determined specific to the research programme or group; hence assessment outcomes are not comparable with other groups or institutions.

This focus on impact and the assessment of non-academic outputs from research fits into a wider picture of increasing need for accountability in research.
SIAMPI team members aimed to be as broad as possible in their application of the approach, selecting case studies covering societal impact in as wide a range of settings, fields, organisations, levels of aggregation, and types of funding as possible. This was partly in response to a desire among the team to expand and develop the concept of Productive Interactions and the previous work of ERiC, but also stemmed from the broad nature of the EC call for proposals to which they were responding when designing the project. To a large extent they achieved this, with case studies spanning different institutional and evaluation contexts in a range of subject areas across four countries (Spaapen and Van Drooge, 2011). These are outlined in Table A16, and described in more detail later, in the section ‘Application to date’.

Evidence from the case studies showed that some elements of the initial design of the SIAMPI team were not applicable in all contexts. For example, in the Netherlands more widely (Spaapen, Dijstelbloem and Wamelink, 2007). The ERiC project itself responded to concern that research assessment in the Netherlands had been too strongly focused on academic quality and had not taken into account the wider societal relevance of research (Smith, 2001). In most countries this has developed into a comparative approach, where differences in impact are used to rate (and hence fund) different institutions comparatively, such as in the REF and RQF approach developed in the UK and Australia respectively. However, in the Netherlands, through ERiC, which now forms the portion of the SEP that evaluates societal relevance, the approach has instead focused on assessment for internal learning and improvement. In fact, SEP has never been used for funding allocation purposes. In the SIAMPI project, this concept is continued and spread to a wider European research community.

---

Table A16
Case studies conducted as part of the SIAMPI study, by country and subject area

<table>
<thead>
<tr>
<th>Health</th>
<th>Nanoscience and technology</th>
<th>ICT</th>
<th>Social sciences and humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>Leiden University Medical Centre, Netherlands Institute for Health Services Research</td>
<td>‘NanoNed’</td>
<td>Department of Informatics, Free University of Amsterdam</td>
</tr>
<tr>
<td>UK</td>
<td></td>
<td>UK e-science programme, Digital Economies Programme</td>
<td>Centre for Business Relationships, Accountability, Sustainability and Society</td>
</tr>
<tr>
<td>France</td>
<td>Institut des NanoSciences de Paris, Laboratoire de Photophysique Moléculaire, Paris, Laboratoire de Physique et de Spectroscopie Electronique, Mulhouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td></td>
<td>Social Sciences and Humanities, Spanish Council for Scientific Research</td>
<td></td>
</tr>
<tr>
<td>EU</td>
<td>FP7 supported ICT work</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Scope**

SIAMPI team members aimed to be as broad as possible in their application of the approach, selecting case studies covering societal impact in as wide a range of settings, fields, organisations, levels of aggregation, and types of funding as possible. This was partly in response to a desire among the team to expand and develop the concept of Productive Interactions and the previous work of ERiC, but also stemmed from the broad nature of the EC call for proposals to which they were responding when designing the project. To a large extent they achieved this, with case studies spanning different institutional and evaluation contexts in a range of subject areas across four countries (Spaapen and Van Drooge, 2011). These are outlined in Table A16, and described in more detail later, in the section ‘Application to date’.

Evidence from the case studies showed that some elements of the initial design of the SIAMPI team were not applicable in all contexts. For example,
the questionnaire they developed to identify Productive Interactions was not appropriate for some of the case studies. This was particularly evident in basic research, where researchers were less likely to be aware of the impact of their work as it was much further downstream, so they tended not to associate the communications they had with stakeholders with Productive Interactions and impact, and indeed did not consider these people as stakeholders, since in many cases they were other researchers. In these cases, the SIAMPI team concluded, either the questionnaire needs to be carefully developed to take this into account, or some other forms of analysis need to be used, such as in-depth research by the evaluation team, and interviews to bring out the relevant interactions through discussion. However, and more reassuringly, the SIAMPI team did find that the broad concept of looking at Productive Interactions seemed to be feasible, with some modifications, across all the case studies, and they believe the approach should be widely applicable.

There are, however, some notable restrictions in the approaches utility. The Productive Interactions approach is specifically designed to assess groups or individuals against their own goals and targets, which inherently places some restrictions on its use. The first is that it is not suitable for making comparisons between different research groups or institutions, as the goals and target will differ between them. Therefore, it is not suitable for the kind of comparative ranking and funding approach used on a national level for funding allocation, for example in the UK or Australia. Nor would it be suitable for similar ranking approaches on a sub-national level. This is difficult not just for funders but also for some institutions to accept, as the concept of being able to benchmark against others is the accepted norm. The team has found that some institutions question the utility of approaches that do not allow them to benchmark their performance. However, it is possible to envisage ways in which the approach could be used for funding allocation, for example by assessing progress against predetermined goals.

Another limitation is in the unit of assessment used. This method needs to be applied to a group or unit which can be said to have common goals against which they can be assessed. Therefore, assessment is best conducted at the levels of smaller units, such as people, groups, research programmes and so on. It is possible to conduct assessments on a broader scale, for example at the institutional or university level. However, goals and the assessment process would necessarily be very different and less focused on the assessment of specific pieces of research.

**Measurement**

The Productive Interactions approach focuses on investigating the mechanisms through which research activities lead to changes in society. Interactions are defined in the work by SIAMPI as taking place between researchers and stakeholders and can take three forms:

- direct personal contacts: any direct communication; might be as simple as a conversation but also include more complex interactions such as research collaborations
- indirect interaction: any type of interaction mediated by a carrier; this could be interaction via a publication of any type, from journal article to clinical guideline, or through the media, an exhibition, film or production, and so on
- financial interaction: where there is some kind of economic exchange between stakeholders and researchers; this could take the form of direct research funding, but could also include IP rights, interaction in relation to contracts, consulting, and so on.

Interactions are ‘productive’ when they lead to efforts by the stakeholders to apply research findings and hence change their behaviour in some way (Spaapen et al., 2011). This is clarified by the examples in Table A17, taken from the SIAMPI project’s final report. (Spaapen et al., 2011).

Methods and tools used to identify and evaluate these Productive Interactions can vary significantly, as shown in Table A17, depending on the types of interaction and the stakeholders concerned. However, one element of the process that remains common across applications is the way in which the assessment criteria are developed. One of the key ideas in the SIAMPI project (and ERiC) is that
research should be assessed against the goals of the researcher, group or institution that is conducting it. Thus different research will be conducted with different goals and purposes in mind, and a one-size-fits-all approach does not produce the best results in learning and development, as it does not take into account the context in which research is conducted.

Although the approach is not methodology-specific, in their final report the SIAMPI team members outline a series of specific tools they have developed in order to deploy the Productive Interactions approach in different circumstances (Spaapen et al., 2011). They describe a bibliometrics tool, termed Contextual Response Analysis (CRA). This is a data mining methodology, which explores internet usage of particular documents and publications. They also outline the way in which stakeholder interviews can be used to identify Productive Interactions and the results that they have led to. However, one of the most developed methods of implementing the Productive Interactions approach is for university audits, largely because here they can build on the well-developed ERiC methodology. This consists of a four-step process, the first three of which are presented in a self-evaluation report for the research group, followed by a four-step process which concerns the assessment of the evaluation committee, as follows:

- Step 1: defining the research group’s mission and objectives; should also indicate how the group intends to achieve this mission, for example through planned outputs, types of research, and existing stakeholder partnerships
- Step 2: description of the productive interactions of the group over the research period
- Step 3: a list of indicators of societal relevance; should reflect the group’s mission and context
- Step 4: assessment of the societal relevance of the research by the peer committee; could include stakeholder viewpoints, for example through including relevant experts on the review committee, or a round table with a survey of stakeholders by the peer committee.

It is easy to see how this process might fit into a wider evaluation also covering academic research outputs.

The data collected through this Productive Interactions approach tends to be new and unique because of the way in which indicators and goals are developed in collaboration with the researchers, reflecting the individual goals of the group or programme. Generally the data are analysed in order to improve and learn, so aggregation is not generally considered to be a necessary step in the process. In this context, there are no ‘results’ of the assessment process per se; rather the outcome is learning and possible improvements in the research and translation process at the institution or for the group. However, it is possible to imagine how the progress of particular researchers or groups in meeting their goals longitudinally could be measured and used as a form of performance metric. These goals would probably need to be agreed in advance with research funders, particularly if this measure of performance is going to be used to determine funding allocation. Performance measurement can also form part of a learning and development process. In the SEP, there is a requirement for ratings on a five-point scale to be produced, from unsatisfactory to excellent. There is provision in ERiC to do this, which could be applied in SIAMPI, comparing results not just against goals, but also considering whether the goals were sufficiently ambitious to warrant a high rating.

<table>
<thead>
<tr>
<th>Productive Interaction</th>
<th>Social impact</th>
<th>Stakeholder</th>
<th>Assessment tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct, personal</td>
<td>Behavioural change</td>
<td>One-to-one, personal and professional networks</td>
<td>Interviews, focus groups</td>
</tr>
<tr>
<td>Indirect, media</td>
<td>Uptake, use</td>
<td>Different audiences</td>
<td>Quantitative data collection</td>
</tr>
<tr>
<td>Financial or in kind support</td>
<td>Collaboration</td>
<td>Joint projects</td>
<td>Annual reports, other documents</td>
</tr>
</tbody>
</table>
Application to date

So far the Productive Interactions approach has only been used in the case study trials conducted in the SIAMPI study. However, these case studies cover a broad range of different research settings. Case studies were conducted in four countries: Spain, the UK, France and the Netherlands. Furthermore, four different research areas are covered: nanotechnology, ICT, health, and social and human sciences. The case studies are described in more detail in the project final report (Spaapen et al., 2011). In the social sciences and humanities, the case studies covered the impact of the Centre for Business Relationships, Accountability, Sustainability and Society (BRASS), a research initiative funded by the UK Economic Social and Research Council (ESRC), and the research groups in social sciences and humanities of the Spanish Council for Scientific Research (CSIC). Approaches to evaluation and funding differ significantly between the UK and Spain. In the UK, impact assessment is well established and part of the funding process, and research funding is awarded over time-limited periods for individual research projects. By contrast, in Spain, there is no formalised approach to impact assessment and CSIC is largely supported by core government funding and supports long-term projects. In these different contexts, the needs of impact assessment differ significantly, so this was a useful test of the applicability of the Productive Interactions approach. The SIAMPI team found that in both cases the Productive Interactions approach was well received and helped the research managers to identify interactions and contributions of which they were not aware. In fact, CSIC intends to use the SIAMPI approach in order to assess a large proportion of its research in the future. BRASS found the approach a useful and credible way to present its engagement activity to research funders.

In health research, the case studies covered two institutes: the Leiden University Medical Centre (LUMC), a large centre which provides care and education along with a wide ranging research programme, and the Netherlands Institute for Health Services Research (NIVEL), an independent non-profit research organisation, which conducts policy-relevant research into healthcare needs, provision and policy in the Netherlands and internationally. Clearly the goals, scale and research types differ significantly between these two centres, and it was found the Productive Interactions approach could be conducted successfully in both cases. The organisations considered using the approach for the assessment of their impact in the future. Although LUMC appreciated the benefits of looking at Productive Interactions, its managers felt they needed an approach which allowed benchmarking and comparison with other institutions. NIVEL managers have decided to pursue the Productive Interactions approach in their future evaluations.

In nanoscience and technology, four organisations were covered by the case studies: the Institute des Nanosciences de Paris (INSP), which consists primarily of a national research centre, but also involves many of the researchers of the University of Paris VI-Marie Curie; the molecular nanoscience research groups at the Laboratoire de Photophysique Moléculaire LPPM (University Paris XI-Orsay) and the Laboratoire de Physique et de Spectroscopie Electronique in Mulhouse; and the research being conducted through the national consortium (‘NanoNed’), focusing on groups from the University of Twente and the Eindhoven Faculty of Electrical Engineering. Nanoscience was chosen as a case study area specifically to test the applicability of the Productive Interactions approach to basic research. In fact, nanoscience is a particularly interesting example, since although the research is basic, work in the field is often touted as having potentially significant implications. This means it might prove a challenge for the Productive Interactions approach to link up these high long-term expectations with current interactions. To some extent, that was found to be the case, if social impact is understood as product use by end-users. However, if a wider definition is used, and multiple steps of interaction are taken into consideration, the approach was able to find Productive Interactions in this field, although it was more challenging than in more directly applied research areas.

In ICT, the case studies covered research in two countries and at the EU level. In the UK, there were case studies on applied ICT research from two national programmes: the UK e-science programme
Productive Interaction. For example, if a group was producing enabling technologies that allowed other research to have an impact this would be relevant for inclusion in the range of Productive Interactions identified. This insight was also helpful in understanding Productive Interactions for more basic research, such as in some of the nanotechnology and ICT work where there were multiple steps, through different research groups, in the process of research being translated to application or practice.

Analysis

The nature of the evaluation framework means that the Productive Interactions approach should meet its aims. The approach and the indicators used are directly tailored to the goals and mission of the research unit and hence the findings should be matched to these expectations. However, the extent to which it leads to learning and improvement is difficult to say, and has not been measured. Many groups have stated that they found the process useful and, as outlined above, many have also decided to continue using the approach after the conclusion of the SIAMPI case study. However, assessing effectiveness in this case is difficult, at least until second and third round deployments at the same locations have taken place to see if learning and improvement have taken place against the baseline laid down in the first round of evaluation. As yet, there is no clear evidence of changes in policy as a result of the implementation of this framework.

Although the adaptiveness and flexibility of the approach is one of the key strengths of the approach, it is perhaps in some senses one of its main weaknesses. Researchers find this kind of open-ended, non-prescriptive approach difficult to manage, and developing unique sets of indicators (in collaboration with the evaluators) is challenging and requires a lot more thought than many other more prescriptive evaluation processes. Of course, the counter argument is that this time is well spent and is an important part of the learning process. Nevertheless, it is probably a deterrent to wider application.
The SIAMPI team hopes to address this through a recent methodology paper, which helps to explain the process and provide ‘tips and tricks’ for developing indicators (Spaapen and Van Drooge, 2011). Another element of the process, which academics can find challenging, is the lack of comparability. Many researchers are competitive by nature and want to know how they are performing in comparison with others. This can also be important for institutions in a range of contexts, from attracting students to applying for funding. This approach does not provide a benchmarked result, a departure from conventional thinking regarding evaluation which many find difficult to accept.

Another strength of the approach is that it is less susceptible to manipulation than other evaluation methods, which set up fixed indicators ahead of time which researchers can then tailor their work to meet. This can produce perverse incentives, which may be damaging in the way they alter research processes. As the indicators are directly tailored to the needs and goals of the research group, these kind of perverse incentives are avoided.

The approach is also better able to reflect the achievements of early career researchers than many other approaches, as evidence of Productive Interactions can be produced as soon as research is conducted, whereas actual societal benefit or impact can take much longer to occur. Also, since the assessment is against the aims and self-image of the research unit, rather than benchmarked against others, the approach is inherently more flexible, and can allow for the different circumstances of early career researchers, those taking maternity leave or leave due to ill health, and others.

As described previously, although results are not comparable with other research units, they may be comparable over time within the same unit provided the goals of the unit do not change significantly, and this could be a method for external assessors to make use of the approach, for example in assessment of unit performance, and potentially even in assessment of funding allocation.

It is not clear to what extent this approach will be used in the future. The SIAMPI project is now coming to a close, so no further case studies are currently planned. A few of the organisations that were part of the original case studies have expressed an interest in continuing to use the approach as described above. However, whether they will actually do so remains to be seen. There have been no clear intentions to use the Productive Interaction approach elsewhere, but the findings of the SIAMPI study are only just being published at this stage, so it is as yet uncertain whether this will be taken up more widely. There is no explicit plan to drive wider adoption of the approach.

**Wider applicability**

As described previously, the Productive Interactions approach can theoretically be applied in a very wide range of contexts because of its flexibility. The challenge is in implementation. This is partly because results are not comparable with other groups or institutions, which is unacceptable for many applications and users. An equally significant challenge is the practicalities faced by groups developing their own indicators. This is intellectually challenging and potentially time consuming, in part because this is a radical departure from normal ways of thinking about evaluation. As yet, there are no concrete estimates for the financial or time costs of implementation. The SIAMPI team is working on materials to help with implementation, such as practical guides with ‘tips and tricks’ on how to perform a Productive Interactions evaluation, particularly in developing indicators.

The critical response to this approach is so far fairly limited, mostly because the findings of the SIAMPI project are as yet largely unpublished. However, the general response from those involved, for example in case studies, has been positive with some reservations corresponding to the implementation challenges described above. It is generally recognised that there are clear advantages in the incorruptibility of the approach and the focus and input to strategy that it provides.\(^{19}\) However, the challenge remains in having the ability and time to develop indicators and think about impact and interactions effectively within institutions.

\(^{19}\) SIAMPI project member, at interview.
There is some potential for this approach to be used in different ways than those originally intended. For example, it is possible to imagine institutions using this approach internally to assess their impact while also using it to produce impact case studies and statements for evaluation processes such as the REF in parallel, but as yet there has been no such application.

**SWOT analysis**

*Figure A16*

A SWOT analysis for Productive Interactions

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formative</td>
<td>High burden</td>
</tr>
<tr>
<td>Sensitive to institutional goals</td>
<td>Not comparable (between institutions)</td>
</tr>
<tr>
<td>Avoids perverse incentives</td>
<td>Challenging to implement</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>Requires assessors to identify productive interactions</td>
</tr>
<tr>
<td>Flexible</td>
<td>Assumptions interactions are a good indicator of impact</td>
</tr>
<tr>
<td>Some tools and ‘how to’ guides being developed</td>
<td>Avoids time lag interaction to impact</td>
</tr>
<tr>
<td>Avoids time lag reduces bias against early career researchers</td>
<td>Avoiding time lag reduces bias against early career researchers</td>
</tr>
<tr>
<td>Multi-disciplinary</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piloted in a range of countries and disciplines</td>
<td>Scalability</td>
</tr>
<tr>
<td>Could support strategic thinking about impact</td>
<td>No implementing owner</td>
</tr>
<tr>
<td></td>
<td>Needs to move from research to operationalisation</td>
</tr>
<tr>
<td></td>
<td>No developing owner – what will happen now FP7 grant funding has run out?</td>
</tr>
</tbody>
</table>
STAR METRICS

Brief description
The STAR METRICS (Science and Technology for America’s Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science) approach is a collaboration between the Office of Science and Technology Policy (OSTP), National Institute of Health (NIH), National Science Foundation (NSF), Department of Energy (DOE), US Department of Agriculture (USDA) and Environmental Protection Agency (EPA) to develop new approaches to assessing performance on investments in science research in the US. It aims to improve policy making, provide accountability and engage the public. It is a metrics-based approach and engages existing data sets, which should minimise burden on researchers and create a comparable and reproducible set of data on performance (Lane, 2010; Lane and Bertuzzi, 2011).

A pilot was conducted in 2009 and a two-step approach developed:

- Level 1 develops measures of the economic support of the science workforce by describing who is employed through the funding provided.
- Level 2 examines the research conducted and develops wider indicators covering the impacts of research on economic growth, workforce outcomes, scientific knowledge and social outcomes.

The approach is impact focused. Although Level 2 includes scientific knowledge creation as an assessment criterion, it is only as one element of a broader package of metrics designed to measure what it means to fund ‘good science’, including wider impacts.

Origin and rationale
The STAR METRICS approach was originally developed through a pilot project by the NIH, NSF and OSTP in collaboration with the Federal Demonstration Partnership (FDP). Following the success of this pilot, the approach has become a wider collaboration including further science research funding agencies.

Two key factors contributed to the development of the STAR METRICS approach. The first was the reporting requirements of the American Recovery and Reinvestment Act of 2009 (ARRA), which required recipients to provide American taxpayers with evidence on the value of the money invested. The second was a direct request from the Office of Management and Budget and OSTP for outcome-oriented goals to be generated for science and technology investment. STAR METRICS was designed to meet these requirements while minimising the burden on researchers. It has two main aims. First, it aims to provide a mechanism by which to generate a reliable and accurate record of the employees supported by R&D funds across institutions. Second, at Level 2 it aims to describe the research that is being conducted and interactions between scientists, and measure the impact of the science conducted on economic outcomes, knowledge generation, and wider social and health outcomes (STAR METRICS, 2010). These dual aims are reflected by the two levels of the approach as detailed in the following sections.

The STAR METRICS approach is a collaboration between the US science funding agencies to build a data platform that supports accountability and provides evidence to improve performance and to support decisionmaking at the policy level. The
metrics approach was selected with the aim of using existing data sets to minimise the burden of extra data collection on institutions and researchers. One of its conceptual predecessors is the Lattes Database, a system used in Brazil, which provides data on 1.6 million researchers across 4,000 institutions (Da Silva et al., 2012). Lattes was developed through a virtual community; researchers were encouraged to use Lattes by making references to the data when making funding and promotion decisions. A unique identification system was also incorporated creating a clean and extensive record of researchers. STAR METRICS also draws somewhat on its direct predecessor, PART, which also used a metrics approach to assess science funding performance, but this was a self-assessment approach, asking programmes to assess themselves against their own strategic (impact) goals through a questionnaire.

**Scope**

In the STAR METRICS approach, analysis is conducted using individual researchers as the unit of assessment, on the basis that ‘science is done by scientists’ (Bertuzzi, 2010). At present, the types of impact covered are limited since Level 1 only describes the jobs that are supported through the various science funding sources reported by research institutions. The analysis conducted under Level 2 seeks to connect funded science with some of the outcomes it produces. The exact metrics that will be used to measure impact in these different areas are still being developed, but it has been indicated that these will include economic, social and health impacts as well as knowledge creation. The extent of application is intended to be broad, potentially spanning all government funded science research organisations in the US. However, as yet, the coverage is significantly lower than this, and it is unlikely to achieve 100 per cent coverage imminently since participation is voluntary. Reports can be generated as frequently as the data on which they are based are updated, and since analysis does not require significant input from research institutions, there is relatively low burden in submitting additional information. The intention is that quarterly reports will be produced, matching ARRA reporting requirements.

**Measurement**

At the current stage of development, STAR METRICS collects data on how many scientists, including graduate and undergraduate students, as well as researchers, are supported by federal science funding. This includes workforce support through sub-awards, sub-contracts and overheads. The nature of science funding in the US means has prevented a central record of this data across institutions. One of the key principles of the STAR METRICS approach is to use the existing administrative systems of each institution to collect this data. The STAR METRICS team works with the institutions to develop an automated system, which extracts 14 key data elements from the data research institutions held in their payroll and accounting systems, thereby minimising the burden of data collection on the participating institutions.

The process for institutions participating in the Level 1 programme takes the form outlined in Figure A17, and consists of six steps. In Step 1, representatives from the participating institution meet STAR METRICS representatives to introduce the programme to institutional authorities. Their decision to join STAR METRICS and the submission of the participation agreement, forms Step 2. Next, the research institution sends sample data drawn from their administrative systems (Step 3), which the STAR METRICS team analyses and on which it provides feedback and reports (Step 4). This allows institutions to hone the data submitted to better match the formatting requirements. Once any such formatting issues are resolved, the institution submits recent and historical data (Step 5), which the STAR metrics team uses to generate quarterly reports and maps in a standardised format (Step 6). After this initial interaction, institutions submit data for ongoing quarterly reporting.

The data collected in this process fall into five categories, as outlined in Table A18.

Significantly greater and more complex analysis is anticipated at Level 2. A range of different approaches has been proposed to describe and analyse the many and complex outcomes that result from federally funded science. Initially, Level 2 will investigate the ‘what’ of science conducted in the US, to
complement the ‘who’ obtained through the Level 1 data collection. Two mechanisms are proposed for this. The first is a combination of data mining and natural language processing to map the research conducted and link it back to particular researchers and research grants. This approach is still under development and could also allow mapping of research trends in the private sector and interactions between fields and sectors. The second proposed mechanism is the creation of an online network called SciENCV through which any scientist can upload a profile to serve a number of purposes. It would provide a system for networking and act as an interface between scientist and the government. Publications would be suggested (which the researcher can then verify), and there would be space for additional information on research goals and interests. Researchers would also be assigned a unique ID, which would allow government bodies to identify them uniquely, eliminating the misidentification that can occur in some bibliometric approaches. The profiles may be a labour saving tool for researchers and institutions with information pulled directly into research funding applications, progress reports and other documentation, which will spare researchers from having to complete forms repeatedly. The approach draws significantly on the Lattes Database system established in Brazil, with the team collaborating and communicating closely with agencies in Brazil, as well as leveraging a variety of existing tools and data in the US, such as the automated data systems of research institutions, several NIH investments (e.g. NIH RePORT – RePORTER and IMPAC II), a variety of NSF investments (e.g. Research.gov and CITESEER), EPA’s HERO database, DOE’s science.gov and publication datasets.

The strategy for the STAR METRICS team in developing Level 2 is to use a staged approach. In the short term (~3 years), its core focus will be on developing the web profile system, developing a core information platform and drawing together existing datasets and databases from a range of sources. In the medium term (3–6 years), it intends to augment and develop the inputs to this profile system and improve automation of the system as well as developing approaches which can be used to gen-
# Table A18
Data elements requested at Level 1 of STAR METRICS

<table>
<thead>
<tr>
<th>Data element</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information on awards</td>
<td><strong>Unique award number</strong> The unique federal number as defined by concatenating the six-position Catalog of Federal Domestic Assistance (CFDA) code with the assigned award ID from the awarding federal agency (such as the federal grant number, federal contract number or the federal loan number). Format example: 47.074 AGS-0120950, where ‘47.074’ is the CFDA and ‘AGS-0120950’ is the federal award ID.</td>
</tr>
<tr>
<td>Recipient account number</td>
<td>Research institution’s internal number for the award.</td>
</tr>
<tr>
<td>Overhead charged</td>
<td>Actual overhead dollars disbursed for award in the specified period.</td>
</tr>
<tr>
<td>Information on individuals</td>
<td><strong>Recipient account number</strong> Research institution’s internal number for the award.</td>
</tr>
<tr>
<td><strong>Unique award number</strong></td>
<td>The unique federal number as defined by concatenating the six-position CFDA code with the assigned award ID from the awarding federal agency (such as the federal grant number, federal contract number or the federal loan number). Format example: 47.074 AGS-0120950, where ‘47.074’ is the CFDA and ‘AGS-0120950’ is the federal award ID.</td>
</tr>
<tr>
<td><strong>De-identified employee ID</strong></td>
<td>Unique employee ID (not social security number) of grant funded personnel.</td>
</tr>
<tr>
<td><strong>Occupational classification</strong></td>
<td>Occupational classification or job description of the funded personnel (ex-faculty, undergraduate student, graduate student, postgraduate researcher, research support, technician or staff scientist, research analyst or coordinator, clinicians).</td>
</tr>
<tr>
<td><strong>FTE status</strong></td>
<td>Designation of the status of the funded personnel (full time = 1.0, half time = .5).</td>
</tr>
<tr>
<td><strong>Proportion of earnings allocated to award</strong></td>
<td>Calculated portion of the earnings charged by funded personnel to the award in the specified period.</td>
</tr>
<tr>
<td>Information on indirect costs</td>
<td><strong>Overhead salary + fringe to total ratio</strong> Proportion of overhead dollars that goes to pay salaries and fringe benefits. The calculation used is (salary $ + fringe $) ÷ total overhead $.</td>
</tr>
<tr>
<td>Payments to vendors</td>
<td><strong>Unique award number</strong> The unique federal number as defined by concatenating the six-position CFDA code with the assigned award ID from the awarding federal agency (such as the federal grant number, federal contract number or the federal loan number). Format example: 47.074 AGS-0120950, where ‘47.074’ is the CFDA and ‘AGS-0120950’ is the federal award ID.</td>
</tr>
<tr>
<td><strong>Recipient account number</strong></td>
<td>Research institution’s internal number for the award.</td>
</tr>
<tr>
<td><strong>Vendor Data Universal Numbering System (DUNS) number</strong></td>
<td>Vendor DUNS number. The vendor’s nine-digit DUNS number. If DUNS is unavailable then substitute zip code with ‘Z’ prefix so as to distinguish it from the DUNS number.</td>
</tr>
<tr>
<td><strong>Vendor payment amount</strong></td>
<td>The funds charged to the award by the vendor in the specified period.</td>
</tr>
<tr>
<td>Sub-contracts and sub-awards</td>
<td><strong>Unique award number</strong> The unique federal number as defined by concatenating the six-position CFDA code with the assigned award ID from the awarding federal agency (such as the federal grant number, federal contract number or the federal loan number). Format example: 47.074 AGS-0120950, where ‘47.074’ is the CFDA and ‘AGS-0120950’ is the federal award ID.</td>
</tr>
<tr>
<td><strong>Recipient account number</strong></td>
<td>Research institution’s internal number for the award.</td>
</tr>
<tr>
<td><strong>Sub-award recipient DUNS number</strong></td>
<td>The sub-recipient organisation’s nine-digit DUNS number. If DUNS is unavailable then substitute zip code with ‘Z’ prefix so as to distinguish it from the DUNS number.</td>
</tr>
<tr>
<td><strong>Sub-award payment amount</strong></td>
<td>The funds charged to the award by the sub-awardee in the specified period.</td>
</tr>
</tbody>
</table>

SOURCE: STAR METRICS (n.d.)
erate connected data about downstream impacts in a standardised way across participating researchers. Finally, over the long term (6 years+), additional data about interactions with the international scientific community will be introduced. Incorporating international data and standards may also enable additional types of benchmarking.

In the longer term, the team hopes to help researchers describe downstream impacts from scientific research and to obtain measures for this in an automated way. A number of approaches have been considered. For example, it is likely that some existing indicators, such as the numbers of start-up companies formed and patents awarded, can be used, and academic outputs tracked through standard bibliometric approaches (Kramer, 2010). Benefits to society are less easily measured through existing data sets, but the team hopes to develop suitable approaches working collaboratively with the scientific community. In all cases, the basic premise is to extract data and information regarding impacts from existing data sets and records without requiring any significant additional data from the scientists or institutions.

Since this approach is still being developed, it is not yet clear how (or if) data from Level 2 will be aggregated or exactly how results would be used. However, it is suggested that outcomes will be made publicly available to meet accountability requirements, and the information should be used by research institutions to assess and improve their performance. Outcomes are not directly linked to funding, but may have a policy impact if they are used to aid strategic decisionmaking, which could influence how funding is directed (to achieve policy aims, rather than specifically reward success).

In the work undertaken at both levels, visualisation is an important part of making the gathered data useful and intuitive. The Level 1 data has been presented in a number of formats, including a map-based format in which jobs supported by federal science funding are mapped across regions. An example is shown in Figure A18, which shows the kind of data included in the reports to individual

---

20 Taken from a presentation by Stefano Bertuzzi, given on 25 May 2011, Montreal.
institutions. Aggregated data across institutions are included in publicly available reports. An example of how this will be presented is shown in Figure A19.

**Application to date**

At present, the data collected through Level 1 of the process are intended for use at the institutional level and more widely once they have been aggregated. First, the participating institutions receive a quarterly report for their institution, with maps showing where jobs are supported by science funding geographically. It describes who they are supporting, where the enterprises they support are based, and has a range of other information about funding for scientific research at their institutions. This is not distributed more widely by the STAR METRICS team; it is confidential information that each institution controls. Generally, it has been used by participating institutions for advocacy purposes as well as for record keeping and monitoring. Ways in which the material has been used so far include presentations to universities’ boards of trustees to illustrate the impact on local jobs, and similarly in presentations to state legislators.

The aggregated data from institutions are intended to be publicly distributed to demonstrate accountability and to meet government requirements to test whether stimulus money has actually generated jobs. The first report was published late in 2011.

At present, the main audience intended for the data collected are the science research funders at the government level and the Office of Management and Budget (OMB). However, the data are also used to justify investment in science to the public, so in some senses, another audience is the US taxpayer. At this stage, there are no consequences intended for funding allocation. The main aim of the process is to obtain a better understanding of whether jobs

---

21 Taken from a presentation by Stefano Bertuzzi, given on 25 May 2011, Montreal.
have been supported, but wider application after Level 2 is intended to build internal learning and to drive policy, which may in turn influence (though indirectly) funding allocation.

There are currently more than 80 academic institutions at various degrees of participation in STAR METRICS, but as it is a voluntary process it is unlikely to achieve universal coverage. Also, although there are around 4,800 universities in the US, only around 120 of these fall into the group of Tier 1 research universities, which conduct nearly all federally funded research. These are the primary targets for the STAR METRICS team, which estimates that with the universities currently committed to the scheme, they cover around 40 per cent of all federally funded university research in the US. However, it is perhaps important to note that this is likely to be a biased sample, given participation is voluntary and therefore likely to favour particular types of institutions, which stand to gain most from participation in this type of programme.

**Analysis**

It is difficult to assess how effective this approach is in achieving its aims as implementation is still ongoing, but feedback from participating institutions has generally been positive so far, and participation levels continue to grow. Level 1, however, is the more straightforward element of the process. The real challenge will come in the implementation of effective approaches to assessing some of the components of Level 2. It will be interesting to see how effectively these elements of the approach can be employed. Similarly, it is hard to assess what the impact of these elements of the approach will be on the incentives they generate for researchers, and their impact on equality and diversity in academia, until it is clear what form they will take.

One of the key strengths of the approach is its aim to minimise the burden on institutions of the process of participating by automating data extraction processes and building approaches on existing data sets as much as possible. Whether this can be maintained in the assessment of, for example, downstream impact remains to be seen, but it remains the aim of the STAR METRICS team.

Another feature of the approach that could be considered both a strength and a weakness, depending on one’s perspective, is that of voluntary participation. This is likely to result in incomplete coverage, and – by extension – a biased sample of research institutions. However, it also ensures that participants are engaged with the process and helps to build a collaborative atmosphere. There are currently no plans to make participation in the scheme compulsory. It is interesting to consider the risks and potential rewards for organisations considering participation in such schemes. On initial consideration, those in high performing institutions might consider that participation would be desirable as they have significant impacts to demonstrate. However, it may also be the case that such institutions, which already have a good reputation, might have the most to lose if they are not able to demonstrate a good return on the comparatively high level of investment made in them. Thus, the decision to participate is not always straightforward, and understanding what biases may be present in the sample of data collected is also complex.

Although participation is voluntary, efforts are being made to encourage engagement. The team behind STAR METRICS engages in a range of outreach activities with institutions, also drawing on its members’ relationship with the FDP (between central government and universities) and other organisations such as the Association of American Universities (a non-profit organisation of 61 public and private research universities in the US and Canada), the Association of Public and Land-grant Universities (a non-profit association of public research universities, land-grant institutions and state university systems) and the Council on Governmental Relations (an association of research universities).

**Wider applicability**

The full costs of implementation are not yet well established, or at least not yet published. However, NSF and NIH invested $1 million on the scheme in its first year of operation (NSF, 2010) (the annual budget for these two organisations is around $38 billion). Several further agencies are now involved, and each agency will continue to contribute $500,000 a
year, in a combination of direct investment and provision of resources, including staff.

There are also costs in time commitment for participating institutions. According to the STAR METRICS team, while each institution has reported different time commitments for setting up their systems to pull the data for Level 1, estimates for initial participants were in the range of 20 to 640 hours, with a median of 100 hours. For subsequent participants, the time taken has fallen, with the range for these later participants being between 30 and 100 hours, with a median of 45 hours. After this initial set-up phase the system is automated and the time required each quarter to upload data is minimal (ranging from 0 to 10 hours, with a median of 2.5 hours). However, this is just to produce data for Level 1. As the research questions with Level 2 emerge and research institutions engage with the STAR METRICS platform to do their own research, they may choose to add additional resources to make even better use of the data.

Although no other organisations are using the approach as yet, it seems that it should be highly transferrable, but requires fairly significant centralised data collection and processing so it is probably a better fit for bigger funding organisations. However, the extent to which this is onerous centrally depends what elements of the approach (from either level) are included. It would be possible to conduct a simplified analysis with a more limited range of metrics, tailored to the interests and requirements of the funder. In all cases the limit is the availability of existing data which can easily be extracted. It will be interesting to see how this can be implemented in more difficult areas, such as societal benefits, as Level 2 progresses. The nature of the approach is that most of the effort required is in setting up the data collection approaches, but that then data collection is relatively automated. Therefore, it may be possible to directly transfer data collection methods developed by the STAR METRICS team, reducing these set-up costs.

There has been significant international interest in the approach. Japan in particular has been engaged with it and has launched the Initiative for Development of Science of Science, Technology and Innovation Policy in April 2011, which draws on the STAR METRICS principles. There has also been engagement at the European level. The STAR METRICS team held a workshop in June with the European institutions, and the European Commission is considering using similar approaches for evaluation of some of their activities. The STAR METRICS team is also briefing funding agencies in Brazil and China on the approach (Grants World Blog, 2011).

**SWOT analysis**

*Figure A20*

**A SWOT analysis for STAR METRICS**

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mining approach is relatively novel</td>
<td>Not fully developed and tested</td>
</tr>
<tr>
<td>Low participant burden once set up</td>
<td>High initial burden, and expertise required to establish</td>
</tr>
<tr>
<td>Not a ranking approach – does not produce a single indicator of comparative performance</td>
<td>Approach beyond Level 1 not proven</td>
</tr>
<tr>
<td></td>
<td>Level 2 will depend on quality of data input</td>
</tr>
<tr>
<td></td>
<td>Level 1 focused entirely on jobs for money input (not comprehensive)</td>
</tr>
<tr>
<td></td>
<td>Summative (at present)</td>
</tr>
<tr>
<td></td>
<td>Not a ranking approach – does not produce a single indicator of comparative performance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mining</td>
<td>Non-participation (not compulsory)</td>
</tr>
<tr>
<td>Harmonisation between funders</td>
<td></td>
</tr>
<tr>
<td>ARRA (Phase 1)</td>
<td></td>
</tr>
<tr>
<td>International interest</td>
<td></td>
</tr>
</tbody>
</table>
AERES

Introduction

The Evaluation Agency for Research and Higher Education (AERES) is an independent administrative authority, which was established in 2007. It is tasked with evaluating French research organisations and institutions, research and higher education institutions, scientific cooperation foundations and institutions as well as the French National Research Agency. The assessment aims to provide feedback to universities for improvement of research activities and degree programmes. AERES has four evaluation cycles, which take place over a four-year period (2008–2011, 2009–2012, 2010–2013 and 2011–2014). Each year 25 per cent of all participating institutions and four or five research organisations are evaluated. AERES undertakes:

• evaluation of research activities conducted by the research units of research organisations and institutions
• evaluation of programmes and degrees of higher education institutions
• approval of the staff evaluation procedures of research organisations and institutions (the focus is on evaluation of procedures, not of the staff).

Basic description

AERES evaluations are based on internal and external assessments. The self-evaluation (internal evaluation) aims to provide each evaluated institution or organisation the opportunity to assess its performances. This provides external experts charged with evaluating the institution with clear and explicit information for assessing the institutional strategy for its research and the range of programmes available; the external review is conducted on the basis of the self-evaluation report and an evaluation application submitted by the institution or organisation. AERES also carries out an analysis of the way in which the institution fulfils its missions.

Peer review and bibliometrics are some of the main tools used by AERES. It is nonetheless recognised that while bibliometrics may be well adapted to biological sciences, for example, this is not the case for all disciplines. Other indicators such as participation in major international conferences are taken into consideration to reflect different practices across disciplines in capturing impact.

AERES uses an integrated evaluation process across research units, programmes and degrees, and higher education and research institutions. This process is organised into three successive stages: the evaluation of research units, followed by the evaluation of degrees (bachelor’s, master’s and doctoral schools), and the evaluation of institutions, which incorporates the results of the first two stages.

At the research unit level, the evaluation assesses a unit’s strengths and weaknesses and enables it to position itself better within its scientific and geographical environment. The evaluation of research units, in addition to the qualitative assessment of self-evaluations, particularly with regard to strengths, weakness (in a SWOT analysis) and recommendations, is based on a scoring system (A+, A, B and C). This scoring system reflects:

• productivity (quality, quantity, impact of scientific research)
• socio-cultural and economic impact (national and international relevance of the research)
Table A19
Assessment of the current state of the AERES research unit

<table>
<thead>
<tr>
<th>Score</th>
<th>Productivity</th>
<th>Socio-cultural and economic impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>International leadership</td>
<td>International appeal</td>
</tr>
<tr>
<td>A</td>
<td>International recognition</td>
<td>National appeal</td>
</tr>
<tr>
<td>B</td>
<td>Significant national impact</td>
<td>Limited appeal (regional appeal)</td>
</tr>
<tr>
<td>C</td>
<td>Requires significant improvement</td>
<td>Insufficient impact</td>
</tr>
</tbody>
</table>

Table A20
Prospects and opportunities for developing research excellence of the AERES research unit

<table>
<thead>
<tr>
<th>Score</th>
<th>Strategy and research life</th>
<th>‘Project’ performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>Excellent environment for nurturing and promoting research excellence</td>
<td>Excellent ‘project’ that promotes cutting-edge research</td>
</tr>
<tr>
<td>A</td>
<td>Largely conducive for nurturing and promoting research excellence</td>
<td>Highly competitive ‘project’ but requires greater clarity and research focus</td>
</tr>
<tr>
<td>B</td>
<td>Improvements on a number of axes required for effective development of a conducive research environment</td>
<td>Competitive at the national level but requiring significant improvement</td>
</tr>
<tr>
<td>C</td>
<td>Significant attention required to create a competitive research environment</td>
<td>Weak feasibility or relevance</td>
</tr>
</tbody>
</table>

- strategy and research life (management, development of research teams and so on)
- the ‘project’ performance (originality, quality, prospects and so on).

The first two elements relate to the current state of the research unit and aim to provide feedback for improvement (Table A19). The other two criteria mainly refer to prospects and opportunities for developing research excellence (Table A20). A general score based primarily on the quality of scientific research is also issued.

The evaluation of programmes and degrees aims to assess the scientific and professional relevance of the range of programmes in knowledge acquisition and skills development. It involves assessing the quality of the range of programmes with regard to the institution’s policy. The evaluation of programmes comprises four scoring levels, A+, A, B and C, and a multi-criterion score for doctoral schools. The institution’s self-evaluation of its programmes is a key element for the external review and is expected to conform to European standards. The assessment also takes into account the relevance of the teaching methods practised and the performance of each programme based on what becomes of graduates for each degree (details of the evaluation process and criteria for each degree level are outlined in the guidelines).

At the institutional level the evaluation analyses the institution’s governance and policy, covering research, exploitation of research findings, training, student life and external relations.

Each of these three evaluations has a preparative step that includes the creation of an internal expert committee. A site visit is organised for institutions, research units and doctoral schools. The process is completed after the post-evaluation step, which includes provision of feedback of the analysis to the institution.

Site visits involve a range of interviews at the institutional level (the institution manager, the management team, the deans, students, professors and staff), research unit level (unit director, unit teams, unit advisory board) and doctoral school level (doctoral school directors, doctoral student representative and board members).
Background

The AERES applies the Standards and Guidelines for Quality Assurance in the European Higher Education Area adopted in Bergen in 2005 by the ministers of higher education in the member countries of the Bologna process. AERES is committed to:

- implement quality assurance according to a process-based approach, which is adapted to the purposes of AERES’ actions throughout its organisational structure and activities
- allocate the necessary resources to this system
- provide a durable framework in order to set and review its quality objectives, to evaluate regularly their adequacy in relation to the needs of the different stakeholders, and to implement the required changes and improvements
- continuously improve the efficiency of its methods and procedures.
The Congressionally Directed Medical Research Program

Introduction

The Congressionally Directed Medical Research Programs (CDMRP) manages biomedical research (cancer research, military medical research and other disease-specific research) for the US Army Medical Research and Material Command (USAMRMC). It was created in 1993 when Congress, in response to grassroots lobbying efforts by the breast cancer consumer advocacy community, tasked the Army with developing and managing an innovative breast cancer research programme. The CDMRP depends on annual congressional appropriations. Assessing productivity and outcomes is a key priority of the assessment of CDMRP-funded projects. The evaluation is primarily aimed at monitoring progress, identifying research outcomes and identifying scientific gaps of the programme’s portfolio.

Basic description

The CDMRP identifies and funds research through a competitive grant system to advance prevention, treatment interventions, quality of life, and the eradication of disease. The Program Evaluation Steering Committee (PESC) within the programme evaluation division meets monthly to design and monitor progress on evaluation projects. PESC assesses research relevance, productivity and accomplishments. The evaluation of projects within specific programmes is carried out by subcommittees. Final reports are delivered to the PESC once a project has been completed.

The progress of every grant and contract is monitored annually for the duration of the award. Funded investigators are required to submit annual and final progress reports that summarise their research efforts and the accomplishments of the project. All reports are technically and contractually reviewed. During each review, outcomes generated from the funded research are identified and classified according to a taxonomy system, developed by the programme evaluation steering committee. This taxonomy system classifies each research outcome by type, tracks the phase(s) of development supported by CDMRP funding, and groups research outcomes into families. Each piece of the taxonomy system (type, phase of development, and family) is captured for each new research outcome and updated for each previously identified research outcome. The taxonomy system not only identifies the outcomes of CDMRP-funded research but tracks its progress along the pipeline from initial discovery through clinical validation throughout the life of each award, so some awards may have multiple outcomes and phases. The CDMRP’s portfolio of research outcomes by type and phase is illustrated in Table A21.

Research outcomes of CDMRP-funded projects are compiled using an innovative classification scheme, an electronic taxonomy coding system, which captures tangible products of funded research for the entire CDMRP portfolio. It identifies outcomes of CDMRP research and helps evaluate the return on investment. The system is currently being used to catalogue and track research advances attributed to CDMRP investigators. The coding system allows staff to assist researchers better with their grants and clinical protocols. Such improvements in grants and information management aim to contribute to disease prevention, treatment and management. Table A22 provides an overview.
Table A21
CDMRP taxonomy system of research outcomes

<table>
<thead>
<tr>
<th>Type</th>
<th>Phase of development</th>
<th>Family (selected examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological molecule</td>
<td>discovery</td>
<td>biomarkers</td>
</tr>
<tr>
<td>Drug</td>
<td>development</td>
<td>cell lines</td>
</tr>
<tr>
<td>Clinical or public health assessment</td>
<td>animal validation</td>
<td>risk factors and assessments</td>
</tr>
<tr>
<td>Clinical or public health intervention</td>
<td>human validation</td>
<td>pharmacologic and therapeutic interventions</td>
</tr>
<tr>
<td>Animal model</td>
<td>phase 0 clinical trials</td>
<td>vaccines</td>
</tr>
<tr>
<td>Biological resource</td>
<td>phase I clinical trials</td>
<td>behavioural interventions</td>
</tr>
<tr>
<td>Data resource</td>
<td>phase II clinical trials</td>
<td>statistical methods</td>
</tr>
<tr>
<td>Methodological resource</td>
<td>phase III clinical trials</td>
<td>detection and diagnostic tools</td>
</tr>
</tbody>
</table>

Table A22
Overview of the CDMRP coding system

Product type
Animal model – non-human animal system, such as a knockout mouse model, that mimics specific biological processes
Biological molecule – human molecular substance, such as a gene, hormone or protein
Biological resource – biological material such as a cell line, used for research purposes
Clinical or public health assessment – potential or tested biological procedure, such as biomarker assays and risk assessments
Clinical or public health intervention – potential or tested medical and/or behavioural procedure, such as a surgical technique or diet modification programme

Stage of development
Discovery and/or development – initial product design, identification, and/or synthesis, or product development and/or testing in vitro systems including cell lines to determine product characteristics
Animal validation – assessment of product characteristics and effects in non-human animal models
Human validation – preclinical assessment of product characteristics and effects in human subjects

Family
Animal models
Biomarkers
Detection and diagnostic tools
Military health and readiness
Pharmacologic and therapeutic interventions
The National Commission for University Evaluation and Accreditation

**Introduction**

The National Commission for University Evaluation and Accreditation (CONEAU) is an autonomous, public body in Argentina, which sits under the Ministry of Education. CONEAU is responsible for the evaluation of public and private universities and accreditation of their undergraduate and graduate degrees. It is also responsible for making recommendations about new state university education projects. Created in 1996 through a requirement of the Federal Higher Education Law, CONEAU is an autonomous organisation. It is overseen by a 12-person body, which is nominated by the legislature, the executive, the national council of university presidents and the academy of education. The staff of CONEAU are trained in different disciplines and are characterised as having very specific knowledge gathered in post about university evaluation processes. CONEAU is funded in full by the National Treasury of Argentina and is included in the national budget.

**Basic description**

CONEAU has three main responsibilities in the higher education sector: accreditation of public and private universities programmes, including approval of foreign educational institutions; institutional evaluations of public and private universities; and making recommendations on new state university projects and provisional authorisation of new private universities (World Bank, n.d.).

The federal law which created and empowers CONEAU specifies that evaluations and accreditations of universities must be carried out with scholars who either participate in peer review committees or on advisory commissions. The law also specifies some key features of the evaluation and accreditation process. For example, accreditation is carried out on the basis of standards determined by a council of presidents from national and private universities, and from the Ministry of Education. CONEAU issues a ‘yes’ or ‘no’ verdict on each accreditation based on these academic standards. However, a ‘no’ decision does not necessarily mean that the programme is discontinued; it usually just triggers a review process within the Ministry of Education, a result of which may be that programmes lose their ability to award valid diplomas.

Evaluation of universities is both an internal and external process. The external evaluation of institutions is carried out every six years and helps to inform the accreditation process. External evaluation is performed using self-assessments conducted by universities on the basis of objectives they autonomously determine. This is an important point – there are no standardised sets of criteria or objectives which apply across the country. However, though CONEAU is not responsible for developing evaluation criteria, it is responsible for developing tools, which are used to support the assessment process. The main goal of the external evaluation is to generate recommendations for improvement, and all evaluations have a strong orientation in this direction (World Bank, n.d.). There is no ranking of institutions or programmes.

All external evaluation reports are published on the CONEAU website.
Background

Argentina was the first country in Latin America to begin delivering graduate programmes. CONEAU and accompanying educational system reforms set in place by the Federal Higher Education Law were established against a backdrop of a rapidly expanding educational system and increasing numbers of students enrolling in university. In the early 1990s, external audits by the World Bank and the Inter-American Development Bank identified problems with the education sector in Argentina. CONEAU was established as there were no quality control mechanisms in place to ensure minimum levels of quality.

CONEAU participates in the Iberoamerican Network for the Accreditation and Quality of Higher Education (RIACES).
Introduction

The National Institution for Academic Degrees and University Evaluation (NIAD-UE) conducts evaluations of education and research activities at Japanese higher education institutions. It was established in 1992 as a degree-awarding institution and in 2000 expanded its role to include the evaluation of universities. NIAD-UE undertakes evaluations to provide feedback to institutions on improving education and research, developing research areas, and fulfilling accountability to stakeholders and the public on their status as public organisations.

Basic description

NIAD-UE is an evaluation and accreditation organisation for universities, junior colleges, colleges of technology, professional graduate schools (law schools). NIAD-UE has two main activities:

- to provide alternative routes to earn an academic degree (bachelor’s, master’s and doctoral); it awards degrees to people recognised as having the same academic ability as those who have graduated from a university undergraduate programme or graduate school
- to conduct independent evaluations based on optional evaluation items on request from universities and other institutions.

NIAD-UE’s evaluation process is carried out by its Committee for National University Education and Research Evaluation. The committee is composed of 30 members including presidents, academics and experts such as journalists and economists. The first evaluation took place in 2008. NIAD-UE developed an evaluation framework and evaluation guidelines. The evaluation process is based on document analysis and site visits. Institutions are required to produce a performance report and submit it to NIAD-UE for assessment.

Each university evaluates the levels of education and research in its departments and graduate schools to assess whether they meet stakeholder expectations. NIAD-UE assesses each university’s self-evaluation performance and results and ranks them on a four-level scoring system (ranging from ‘far exceeding expected standards’ to ‘below expected standards’).

The research evaluation is based on a record of research activities and research achievements. Research activities are subjected to quantitative analysis using data provided by the university. The evaluation looks at the extent of research activities and their relationship to the research objectives of the institution. Data on research activities relate to research published in academic papers and books, presentations delivered at academic societies, intellectual property rights applied for and granted, implementation of joint research, and implementation of commissioned research. Access to research funding is also assessed (grants received, external competitive funding, funding for joint research, funding for commissioned research, donations, and donations for endowed chairs).

Research achievements (outcomes of the research activities) are based on a qualitative analysis of the organisation’s research. Evaluation is carried out on academic aspects (publication in scientific journals; reviews by referees submitted publications; critiques, reviews, introductions and references in academic magazines and newspapers; impact factor of academic journals publishing the papers; the papers’
Based on annual plans, mid-term plans and mid-term objectives for education, research and management, NIAD-UE’s evaluation of institutions focuses on the attainment of mid-term objectives for education and research. The first evaluation round of universities, junior colleges and colleges of technology took place in 2005 and in 2007 for law schools. Evaluations are carried out on request by the institutions.

NIADU-UE also engages in research for university evaluation and other institutions. The research aims to improve the evaluation system and in doing provide better evaluation feedback to institutions. An optional evaluation on research conditions of institutions was introduced in 2011.

Peer reviews are conducted on performance levels in 66 fields using an established set of evaluation standards. Each department is required to submit research achievements reported during the four years preceding the year of evaluation.

**Background**

The National University Corporation Evaluation Committee, which is responsible for performance-based evaluation of national university corporations and inter-university research institute corporations, commissions NIAD-UE to carry out evaluations based on annual plans, mid-term plans and mid-term objectives for education, research and management. NIAD-UE’s evaluation of institutions focuses on the attainment of mid-term objectives for education and research. The first evaluation round of universities, junior colleges and colleges of technology took place in 2005 and in 2007 for law schools. Evaluations are carried out on request by the institutions.

NIADU-UE also engages in research for university evaluation and other institutions. The research aims to improve the evaluation system and in doing provide better evaluation feedback to institutions. An optional evaluation on research conditions of institutions was introduced in 2011.
The National Research Foundation – Knowledge Management and Evaluation Directorate

Introduction

The Knowledge Management and Evaluation (KME) Directorate of the National Research Foundation (NRF) monitors and evaluates research, including evaluating and rating individual researchers across the scientific disciplines. A peer review-based, voluntary evaluation system is used, with individual researchers applying directly to the NRF. The NRF national research evaluation system is publicly accessible; individuals can be searched by institution, surname and discipline. An individual’s NRF rating is valid for six years. The rating itself does not preclude individuals from receiving NRF funding. However, after a finite period of time, a researcher must have an NRF rating in order to continue receiving NRF funding (NRF, 2007).

Basic description

The NRF’s peer review-based, voluntary rating and evaluation process is conducted over several stages. It is designed with multiple checks and balances to ensure fairness and consistency (NRF, 2007, pp. 16–18; NRF, 2010, pp. 20–1). Most evaluation is carried out by external international experts who have no monetary incentive for their review. They assess individual researcher applications according to the quality and impact of each applicant’s research outputs and achievements.

There are 23 subject-specific specialist advisory committees, which oversee the rating process, covering disciplines from animal studies to theology. Their role in the evaluation process is to reach consensus and assign ratings on individual researchers based on the peer review, to evaluate the quality of the peer reviewer reports, and to select constructive feedback for the applicant (Marais, 2007). The Executive Evaluation Committee finalises the highest level recommendations (A and P) and the Appeals Committee oversees disagreements.

Researchers apply to each specialist committee with details of their research outputs and achievements over the past eight years, including:

- publication lists, including peer-reviewed publications, grey literature, conference proceedings and patents
- self-assessment statements
- lists of postgraduate students trained
- suggestion of ten national and international peer reviewers (Krige, 2007).

After the application is submitted, the relevant specialist committee identifies a minimum of six peer reviewers for each application. This includes at least three peer reviewers from the list provided by the applicant and three identified by the committee. The peer reviewers evaluate the applicant’s research outputs against ‘internationally benchmarked standards of excellence and impact’. Impact on the field is considered, as well as adjacent fields of research. The committee also estimates the applicant’s standing against other South African and international researchers.

The relevant specialist committee reviews the peer reviewer report and designates each report as ‘excellent’ through to ‘unsatisfactory’, thereby ensuring that each report is validated. Once there are enough ‘satisfactory’ to ‘excellent’ reports, they are passed to an assessor who makes an initial rating determination. Assessors are chosen for their research excellence and will have previously served on a specialist
Appendix C

• the creation and development of the “intellectual infrastructure” of disciplines
• the invention or generation of ideas, images, performances and artefacts where these manifestly embody new or substantially developed insights;
• building on existing knowledge to produce new or substantially improved materials, devices, products, policies or processes’ (NRF, 2010).

Specifically excluded are research activities that include routine testing or analysis of materials which do not involve the development of new analytical techniques, or methods that involve the development of teaching materials or practices which do not involve the use of new knowledge or ideas as defined above.

The research and evaluating rating system has undergone several changes over the years. In 2005 a major review was published and the system was altered for the 2007 review year. On the back of this a further exercise was conducted in 2010. The NRF aims to learn from and improve the system continually.

Committee. The assessor’s initial assessment is discussed and agreed in a final meeting attended by the assessor and the specialist committee. The rating levels are summarised in Table A23.

The NRF system is not formally used to rank universities at a national level, but many use the outcomes of the NRF’s evaluation to self-rate themselves. Some institutions also use the results to provide incentives to staff within their organisations.

Background

South Africa was one of the first countries to pioneer a rating system for individual scientists in the early 1980s under the vision set out by Professor Jack de Wet, Research Grant Advisor at the then Council for Scientific and Industrial Research (CSIR), now the NRF. The system was first applied only to natural scientists and engineers, but since 2002 has included all researchers, including social scientists and humanities researchers.

For the purposes of the rating and evaluation exercise, the NRF defines research in the following way:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
<th>Sub-categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: leading international researcher</td>
<td>International recognition by their peers for the high quality and impact of their research</td>
<td>A1: all reviewers concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2: the majority of reviewers concur</td>
</tr>
<tr>
<td>B: internationally acclaimed researcher</td>
<td>Considerable international recognition by their peers for the high quality and impact of their outputs</td>
<td>B1: all reviewers concur, with some giving an A-rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2: all or majority concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B3: most reviewers concur</td>
</tr>
<tr>
<td>C: established researcher</td>
<td>Sustained recent record of productivity in the field; recognised by peers as having produced a quality body of work with a coherent core</td>
<td>C1: all reviewers concur with giving a B-rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2: all or majority concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C3: most reviewers concur</td>
</tr>
<tr>
<td>P: NRF president’s awardee</td>
<td>Young researchers demonstrating potential through doctoral and existing research; likely to become international leaders</td>
<td></td>
</tr>
<tr>
<td>Y: promising young researcher</td>
<td>Young researchers who are recognised as having the potential to establish themselves as researchers</td>
<td>Y1: all or majority agree there is potential to be an international leader</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y2: all agree there is potential to be an established researcher</td>
</tr>
</tbody>
</table>

Table A23
Rating levels of the NRF

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
<th>Sub-categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: leading international researcher</td>
<td>International recognition by their peers for the high quality and impact of their research</td>
<td>A1: all reviewers concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2: the majority of reviewers concur</td>
</tr>
<tr>
<td>B: internationally acclaimed researcher</td>
<td>Considerable international recognition by their peers for the high quality and impact of their outputs</td>
<td>B1: all reviewers concur, with some giving an A-rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2: all or majority concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B3: most reviewers concur</td>
</tr>
<tr>
<td>C: established researcher</td>
<td>Sustained recent record of productivity in the field; recognised by peers as having produced a quality body of work with a coherent core</td>
<td>C1: all reviewers concur with giving a B-rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2: all or majority concur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C3: most reviewers concur</td>
</tr>
<tr>
<td>P: NRF president’s awardee</td>
<td>Young researchers demonstrating potential through doctoral and existing research; likely to become international leaders</td>
<td></td>
</tr>
<tr>
<td>Y: promising young researcher</td>
<td>Young researchers who are recognised as having the potential to establish themselves as researchers</td>
<td>Y1: all or majority agree there is potential to be an international leader</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y2: all agree there is potential to be an established researcher</td>
</tr>
</tbody>
</table>
The Performance Based Research Fund

Introduction

The Performance Based Research Fund (PBRF) is a tertiary education funding system set up in 2002 to encourage and reward research excellence in New Zealand’s degree-granting institutions. This involves assessing the quality of research carried out by degree-granting tertiary education organisations (TEOs) and funding them on the basis of their research performance. The purpose of the PBRF is not to provide funding for research projects per se, but to reward research excellence and support TEOs to provide an environment that produces research of a high quality. It intends to ensure that degree-level teaching is underpinned by high-quality research activities.

Basic description

The PBRF funding formula is based on three elements:

- quality evaluation: the assessment of the research quality of TEO staff members, based on peer review (60 per cent of the fund)
- a postgraduate research degree completions (RDC) measure: the number of postgraduate research-based degrees completed in the TEO (25 per cent of the fund)
- an external research income (ERI) measure: the amount of income for research purposes received by the TEO from external sources (15 per cent of the fund).

The RDC and ERI measures are calculated annually using weighted three-year averages. Each TEO’s share of funding for each of these three components is determined by its performance relative to other participating TEOs.

The quality of an individual’s research contribution is assessed through the external peer review of their research as presented in an evidence portfolio (EP). The EP forms the basis of the quality evaluation measure and has three key components:

- research outputs (ROs): the outputs of a staff member’s research
- peer esteem (PE): an indication of the quality of the research of the staff member, as recognised by their peers
- contribution to the research environment (CRE): the staff member’s contribution to a vital high-quality research environment, within the TEO and beyond it.

Each eligible staff member has one EP for each PBRF quality evaluation round.

The main aims of the PBRF are to:

- increase the average quality of research
- ensure that research continues to support degree and postgraduate teaching
- ensure that funding is available for postgraduate students and new researchers
- improve the quality of public information on research output
- prevent undue concentration of funding that would undermine research support for all degrees or prevent access to the system by new researchers
- underpin the existing research strength in the tertiary education sector.

Excellence is viewed as going beyond the production of high-quality research articles, books, exhibitions and other forms of research output to include:
• the production and creation of leading-edge knowledge
• the application of that knowledge
• the dissemination of that knowledge to students and the wider community
• supporting current and potential researchers (e.g. postgraduate students) in the creation, application and dissemination of knowledge.

Background
The report Investing in Excellence (PBRF Working Group, 2002), delivered by the PBRF Working Group and endorsed by cabinet, formed the basis for the implementation of the PBRF. Quality evaluation rounds were held in 2003 and 2006, and a third round will be held in 2012. The evaluation strategy has three phases and operates for an approximate ten-year period from mid-2004 to late 2014. Phase 1 focused on the design and implementation of the 2003 quality evaluation. Phase 2 was an independent strategic review of the positive effects and unintended consequences of the PBRF on the sector. Phase 3, the longer-term phase, will focus on whether the PBRF has fulfilled its stated objectives and whether the overall benefits have exceeded the costs. Phase 3 will be undertaken after the 2012 quality evaluation.

The PBRF is guided by the following principles: comprehensiveness, respect for academic traditions, consistency, continuity, differentiation, credibility, efficiency, cost effectiveness, transparency, complementarity and cultural inclusiveness.

The operation of the PBRF involves five major participants:

• TEOs, whose role is to provide complete and accurate data on:
  • census information to determine which staff members are eligible for participation in the 2012 quality evaluation
  • individual staff members’ research activities and contributions during the assessment period in the form of EPs (as part of the quality evaluation)
  • numbers of postgraduate research degree completions (as part of the RDC measure)
  • external research income (as part of the ERI measure)
• peer review panels, expert advisory groups and specialist advisers, whose role is to evaluate the quality of the EPs submitted by the participating TEOs and to assign each of them a quality category
• the TEC Secretariat, whose role is to provide technical, policy and administrative support to the PBRF process and peer review panels and expert advisory groups; in particular, the chairs of those panels and groups
• the Moderation Panel, whose role is to:
  • generate consistency across the peer review panels, while at the same time not reducing the panel judgements to a mechanistic application of the assessment criteria
  • provide an opportunity for independent review of the standards and processes being applied by the panels
  • establish mechanisms and processes by which material differences or apparent inconsistencies in standards and processes can be addressed by the panels
  • advise the TEC Board on any issues regarding consistency of standards across panels
• the TEC Board, whose role is to consider and approve the findings of the quality evaluation for funding and reporting purposes.

22 The Tertiary Education Commission (TEC) commissioned an independent review, which took place in 2008.
Introduction

The Spanish State Programme for University Quality Assessment (PNECU) was established in 1995 and lasted five years (1995–2001). It issued annual calls for institutional programmes for quality assessment in which public and private universities could participate. Although the programme was not compulsory, almost all universities participated.\(^{23}\) The creation of the European Higher Education Area led to the introduction of new regulations aimed at restructuring universities.\(^{24}\) The assessment of universities aimed to provide informed recommendations for improvement. The PNECU evaluation framework was based on the recommendations of the Experimental Programme for Assessment of Quality in the University System, which evaluated teaching, research and institutional management in several universities (its primary purpose was to try out various methods and make proposals for change based on the experiences gained). PNECU also drew on the European Pilot Project for Evaluating Quality in Higher Education, which aimed to test a common methodology among European universities. PNECU sought to develop methodologies that were homogeneous with those in the European Union to establish standards that allowed comparative assessments.

Basic description

PNECU had two types of assessment projects:

- thematic projects related to either one undergraduate degree or to a group of them within the same field of enquiry in one or various universities
- global projects covered by one or various universities considered as a whole.

Both projects assessed quality in teaching, research within the universities, and university services management. Assessment reports were produced for each one:

- Teaching assessment was based on the teaching programmes, including assessment of the programme structure, teaching procedures, student and staff characteristics, and resources and outcomes.
- Research assessment evaluated research in the departments related to the programmes assessed in teaching and covered the department's research objectives, human and material resources, research activity, productivity, quality indicators and research outcomes.
- Assessment of the management of units and services focused on economic and administrative efficiency, decisionmaking procedures, student services structure and facilities in general. It aimed to provide feedback for improvement.

The institutional review process was similar to that in the European context and employed two methods: an internal evaluation (or self-evaluation) and external evaluation.

The internal evaluation was carried out by the evaluation committee of each university. It aimed

\(^{23}\) Spain has 48 public universities in 17 regions and 20 private universities.

\(^{24}\) The accreditation requirement for official programmes was another factor that encouraged universities to participate in the evaluation.
to promote reliable information on the evaluated unit and promote awareness of quality issues in the university community. The external evaluation consisted of a visit by an external committee composed of experts in the field (academic and non-academic). They interviewed leaders, staff and students in each evaluated unit. The external committee compared its findings to those of the internal evaluation, then published a report synthesising the results of the internal and external evaluation.

The reports were required to focus on the analysis, opinions and judgements of those involved in the evaluated units and to make recommendations for improvement. They contained descriptions and context of units evaluated; information on aims and objectives, resources, structure and results; judgements by the evaluation committee about the strong and weak points of the unit; proposals and recommendations for improvement; and relevant quantitative indicators.

**Background**

PNECU was headed by the Council of Universities and a technical committee composed of the council’s officials and evaluation experts was charged with the evaluation process. The technical committee produced written guidelines that aimed to standardise the evaluation process. PNECU had the following objectives, which were largely met:

- promoting institutional assessment of university quality
- providing methodological tools to universities for the assessment process that would be homogeneous throughout the country and similar to processes used elsewhere in Europe
- providing society, especially students, with relevant and reliable information about the quality of the institutions, their programmes, services and scientific levels
- providing accountability to the regional governments (to aid decisionmaking).

PNECU was succeeded by the Plan de Calidad de las Universidades (PCU), which was established in 2001. The general objectives of the PCU are to:

- encourage institutional assessment of higher education quality
- promote and consolidate a legal framework in which regional governments assume a leading role in the management of the PCU in order to encourage regional assessment agencies and establish a network of agencies coordinated by the Council of Universities
- develop homogeneous methodologies for quality assessment integrated in the current practice of the European Union in order to guarantee standardisation of assessments
- provide objective information on the standards reached by each university, which can be used by different bodies as the basis of decisionmaking in their own areas of responsibility.

PCU developed a catalogue of indicators. This is a selection of a limited number of indicators for which definitions, rules for calculation and interpretations are provided. The indicators are classified into supply, demand, human resources, financial resources, physical resources, processes and results. The aim is to provide a common language to share information among institutions that can improve the mechanisms for benchmarking. These benchmarks are used for different purposes, including accreditation.
The Standard Evaluation Protocol

The Standard Evaluation Protocol (SEP) is the approach used to evaluate publicly funded research in the Netherlands. It has existed since 1994, with revision of the protocol taking place in 1998, 2003 and 2009, with the current protocol scheduled to be reviewed in 2015. The key objectives of the SEP are to:

- improve research quality based on an external peer review, including scientific and societal relevance of research, research policy and management
- be accountable to the board of the research organisation, and towards funding agencies, government and society at large

The review consists of a self-evaluation and an external review, including a site visit, every six years. The review is retrospective and prospective, and aims to meet accountability needs and to aid improvement and development. Research institutes (defined as ‘a group of researchers with an articulated shared mission operating under the same management’) in the Netherlands are in the jurisdiction of a university board (represented by the Association of Universities in the Netherlands), the Royal Netherlands Academy of Arts and Sciences or the Netherlands Organisation for Scientific Research. These bodies are responsible for conducting the evaluation, and for developing the SEP.

Evaluation takes place at two levels: the research institute and the specific research groups or programmes. Evaluations look at three main tasks of the research institute:

- production of results relevant to the research community
- production of results relevant to society
- training of PhD students.

The assessment is against four main criteria, which reflect these tasks: quality, productivity, societal relevance, and vitality and feasibility. At the institute level, assessment focuses on policy and strategy, with an emphasis on improvement, while assessment at the programme level focuses on performance and accountability. Correspondingly, at the institute level, assessment is largely qualitative, but at the programme level performance is rated on a five-point scale from unsatisfactory to excellent against each of the four main criteria. The way in which the four criteria are assessed, and how they are broken down into ‘sub-criteria’, is outlined in Table A24.

The required content of the self-evaluation report is clearly laid out in the guidance (VSNU, KNAW and NWO, 2009), and reflects the assessment criteria above. Reports, each of around four or five pages in length, are produced for the institute and for each research programme within it. Standardised approaches are specified for quantitative data compilation, such as success rates of PhD students, to aid comparability. One key element of the self-evaluation report is the SWOT analysis, which is intended to address one of the main evaluation objectives: the improvement of research and research management. This should be transparent, benchmark the institute’s position, and lead to conclusions on future strategy.

The evaluation committee, appointed by the assessing body with input from the institute, usually conducts a site visit consisting of meetings with the director or manager of the institute, its research leaders, a number of tenured and non-tenured staff,
The SEP was last reviewed in 2009. Findings of this review suggested that users valued the evaluation process and the continuity of the approach, but that it should be adapted to reduce administrative burden where possible, and more emphasis should be placed on societal relevance, positioning and benchmarking. These suggestions have been incorporated in this latest version of the protocol, with key steps being reduction in the size of the self-evaluation report and making the mid-term, review light touch.

and a number of PhD students. The committee then produces a final evaluation report (of 15–20 pages) including assessment at institute and programme level. This is made public, along with the final position of the board on evaluation outcomes and recommendations.

Additionally, institutes conduct an internal mid-term review roughly three years after external evaluation to assess progress and formulate actions in preparation for the next review. Timelines and requirements are at the discretion of the assessment board.

Table A24
Assessment criteria, sub-criteria and aspects to be considered for the SEP

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Sub-criteria</th>
<th>Aspects that may be considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>A1 Quality and scientific relevance of the research</td>
<td>Originality of the ideas and research approach, including technological aspects; significance of the contribution to the field; coherence of the programme; quality of the scientific publications; quality of other output; scientific and technological relevance</td>
</tr>
<tr>
<td></td>
<td>A2 Leadership</td>
<td>Leadership of primary individuals; mission and goals; strategy and policy</td>
</tr>
<tr>
<td></td>
<td>A3 Academic reputation</td>
<td>(Inter)national position and recognition; prominence of the programme director and other research staff; impact and significance of research results in the field</td>
</tr>
<tr>
<td></td>
<td>A4 Resources</td>
<td>Human resources; funding policies and earning capacity; relevance of research facilities</td>
</tr>
<tr>
<td></td>
<td>A5. PhD training</td>
<td>Objectives and institutional embedding; structure of programmes; supervision; success rates; educational resources</td>
</tr>
<tr>
<td>Productivity</td>
<td>B1 Productivity strategy</td>
<td>Productivity goals; publication strategy; rewards and sanctions</td>
</tr>
<tr>
<td></td>
<td>B2 Productivity</td>
<td>Scientific publications and PhD theses; professional publications; output for wider audiences; use of research facilities by third parties</td>
</tr>
<tr>
<td>Relevance</td>
<td>C Societal relevance</td>
<td>Societal quality; societal impact; activities aimed at making research results available and suitable for application</td>
</tr>
<tr>
<td>Vitality and feasibility</td>
<td>D1 Strategy</td>
<td>Strategic planning; investments and collaboration; research topics planned for the near future and their perspectives; flexibility and anticipation of expected changes</td>
</tr>
<tr>
<td></td>
<td>D2 SWOT analysis</td>
<td>Analysis of the position of institute and programmes; analysis of strengths and weaknesses</td>
</tr>
<tr>
<td></td>
<td>D3 Robustness and stability</td>
<td>Research facilities; financial resources; staff competition; mobility and attractiveness; expertise within the institute</td>
</tr>
</tbody>
</table>
This appendix reviews the tools used in research evaluation, drawing on those methods found in the frameworks investigated. These are the tools we investigated:

- bibliometrics
- surveys
- logic models
- case studies
- economic analysis
- peer review
- data mining
- interviews
- data visualisation
- site visits
- document review.

These have been identified through the frameworks investigated, and through wider knowledge of the field within the team, and span the key methods used in research evaluation. For each tool, a short report is presented, summarising the way in which it is used, with practical examples for the more complex tools. The majority of the tools are described using the following reporting template:

- **Introduction:**
  - What is it? – define the tool
  - What are the key principles?

- **When should [the tool] be used?:**
  - What are the key limitations of the tool? What are the key factors that indicate it is not appropriate for use in a particular situation?
  - What are its key strengths? What can it add to an analysis? What factors indicate when it is a good or the best choice of approach?
  - At what scale of aggregation can the tool be used? At what level should it be implemented?
  - What is the cost and burden of implementation of the tool?
  - How does it fit with other analysis tools? Are there particular widely used combinations, and what are the reasons for them?
  - How is [the tool] used?:
    - What are the different ways in which this tool can be used and when are they most appropriate?
    - What are the aims of these approaches?
    - How is the tool employed? This section should give a ‘how to’ guide (with caveats) for each core mode of operation.
    - What skills and/or experience are needed?
    - What can go wrong? What are the key challenges in its use?
    - What outcomes can you expect from the tool, and what types of evidence and information are produced?

- **Example case studies:**
  - Give an example of where the tool has been used to illustrate some of the points in the previous sections.
  - Try and draw on a relevant situation in research evaluation where possible.
  - The example should be used to illustrate some of the points made in previous sections.
  - Where a tool has several distinct modes of operation, more than one case study might be useful.
  - Multiple case studies might also be used to illustrate strengths and weaknesses, or the effectiveness of the tool in combination with others.

- **Summary.**
However, for some of the tools, a briefer summary is presented, covering the key features of the approach. This shorter review is provided for those tools where application is relatively straightforward, but where it is useful to set the tool in the research evaluation context, for example document review. The review of tools is based on a review of existing literature, and the expertise and experience of the team.
Introduction

The bibliometrics method involves a number of techniques for assessing the quantity, dissemination and content of publications and patents. In the case of publications, it uses quantitative analysis to measure patterns of scientific publication and citation, typically focusing on journal papers. It is one of a set of evaluation methodologies that may be used to help assess the impact of research. It is used to support research allocations, re-orientations of research support, rationalisation or even wholesale restructuring (Moed, 2007).

Bibliometrics offer insights principally along the following dimensions:

- **activity measurement**: counts of articles as a measure of the volume of outputs in a given research field; in effect, a measure of the size of scientific activity
- **knowledge transfer measurement**: on the basis that the citation process reflects the communication of knowledge within the scientific community and provides an indirect measure of research quality
- **linkage measurement**: involving the assessment of links between individuals and research fields, as an indication of the way in which the social and cognitive networks of scientific research are developed and sustained (mapping exercises)
- **citation analysis**: as a proxy for one or more dimensions of the quality of scientific output.

When should bibliometrics be used?

Some of the strengths and weaknesses of bibliometric approaches relate to:

- changes in the way in which scientific knowledge is disseminated
- precisely what kind of impact these approaches can measure
- precisely what kind of measure of intellectual ‘influence’ bibliometrics truly provides.

There are considerable advantages to a bibliometric approach, especially as the power and range of indicators available improves – but a clear understanding of limitations and caveats is required. From a theoretical perspective, some doubts remain as to the ability of bibliometric methods to capture abstract concepts such as research ‘quality’. Methodological challenges include issues of journal coverage in major bibliometric databases, adequately identifying author affiliations, and choosing the right timeframe for analysis. Caveats to bibliometric analyses include variations in citation behaviour between fields and individuals, and a perennial difficulty in evaluation: attribution. While it is usually possible to determine whether research work contributed to the content of particular publications, attributing publications solely to particular bodies of research is often very difficult if not impossible. These are other limitations of the tool:

- It treats only publications and patents as programme outputs and ignores other outputs and long-term outcomes.
• Time must pass before extensive publication and/or patent citations can be observed.
• Estimates of quality based on citations can be misleading – counts indicate quantity and impact of output, not quality.
• Not all publications and patents are of equal importance – data must be normalised for comparative analysis, for example in the case of publication across research fields and journals.
• Propensities to publish and to patent differ among scientific and technological fields.
• Bibliometrics cannot measure future potential.

Bibliometric approaches offer important advantages over other research evaluation methods. They can be used to generate useful quantitative indicators of collaboration and the nature and extent of interdisciplinary research. As the sophistication of analytical tools improves, they are also being used to develop more general indicators of ‘quality’ and even ‘excellence’. These analyses are supported by a range of indicators of varying complexity that have been developed over recent years.25

The analytical power of bibliometrics derives mainly from use in combination with other methods, rather than independently. There is growing consensus, for instance, that bibliometric analysis can be successfully used to complement peer review decisions. This is particularly the case for large-scale peer review exercises, where the volume of material to be analysed, and indeed its complexity, may be such that some form of quantitative validation may be useful.

Bibliometric analysis has a number of benefits:
• It is widely applicable to evaluation of programmes with an emphasis on publishing or patenting.
• It can address a variety of evaluation topics, including research output, collaborative relationships, and patterns and intensity of knowledge dissemination.
• Diverse audiences can understand the results.
• It can be applied to a programme with a relatively short time lag.
• It has a high degree of credibility and can be used to indicate quality of output.
• It enables analysis of trends including global trends.
• It generates quantitative indicators for research evaluation, which can be very useful in assessing large volumes of material that may have a large degree of complexity.

Bibliometric analysis can be used at a number of levels to support a variety of analyses depending on the size and characteristics of the data set in question. Although initially viewed as a means for building system-level understanding of outputs and impacts in research fields, recent refinements have meant that it is now possible to apply some of the tools of bibliometric analysis to institutions, research groups, and even individuals.

Bibliometrics is used to measure the ‘output’ of individuals, research teams, institutions and countries, to identify national and international networks, and to map the development of new (multi-disciplinary) fields of research.

As a quantitative approach bibliometric analysis is thought to generate significant cost savings and to significantly reduce the administrative burden (Oppenheim, 1997). When compared with alternative methodologies, bibliometrics offers the particular advantage of using apparently objective data that can be collected with minimal involvement from the researchers themselves. This helps to reduce the administrative burden posed by many of the alternatives – particularly peer review – for researchers, while ensuring that data obtained are more likely to be representative because they are gathered for entire research communities, and not simply those researchers who give their consent (Smith, 1981).

How are bibliometrics used?

The integration of bibliometric analysis with other performance analysis tools and in particular peer review is common. Combinations of bibliometric analysis and peer review appear to offer greatest potential in overcoming areas of weakness in the

25 Robust bibliometric analysis requires a clear understanding of the strengths and limitations of each of these measures – and sensitivity to the contexts in which they are used.
peer review system (Moed, 2005). First, the results of bibliometric analysis may be used to challenge peer reviewers, seeking explanations for unusual or unexpected patterns. Second, they may be particularly useful for high-level or multi-factorial analysis during research evaluation exercises – when the number of factors under consideration becomes so high that peer reviewers may have difficulty analysing them all in detail, or where some alternative form of validation of peer review decisions may be required.

Bibliometric analysis is used to measure a range of elements including measuring scientific collaboration, assessing interdisciplinary research and research ‘quality’ and ‘excellence’. It is also used to map scientific research in the following ways:

- **Measuring scientific collaboration:** One of the areas in which bibliometrics offers particularly powerful insights is analysis of collaborative scientific activity – measured typically through the institutional affiliations of authors on publications. This also allows for some comparison of national and international level cooperation – although it should be noted that analysis of collaboration is subject to methodological issues around attribution of research work.

- **Assessing interdisciplinary research:** There are three distinct approaches to assessment of interdisciplinary research. First, a forward-looking, raw publication analysis builds a profile of the contribution of a particular research institute or individual across a number of fields to assess research impact. A measure of multi-disciplinarity can then be derived from the distribution of the group or individual’s papers among fields. A second approach involves looking back at the fields and disciplines of papers that authors draw information from as inputs to interdisciplinary research. A third approach involves looking at the fields into which journal publications produced by an author or research group fall.

- **‘Quality’ and ‘excellence’:** One approach takes into account the skewed nature of citation distributions to provide a measure of the number of publications produced by an institution that are in the top 10 per cent of the worldwide citation distribution of the field concerned. This measure has been used to monitor trends in the position of research institutions and groups at the top levels of their respective fields (Van Leeuwen et al., 2003). Nonetheless, the concepts of ‘quality’ and ‘excellence’ in research evaluation remain fraught with difficulty.

- **Mapping scientific research:** Mapping exercises support a number of high-level analyses such as providing a sense of activity across the sciences, using journal-to-journal citation maps, providing an understanding of core areas of scientific output, and providing an analysis of linkages between fields, domains and sub-domains. Bibliometrics as a mapping tool is also used to support longer-term, strategic decision-making by providing an assessment of changes in the activities of research groups or individual researchers over time and identifying new or emerging fields.

**What can go wrong? What are the key challenges in its use?**

These are some of the theoretical and methodological issues that require attention in using bibliometric analysis:

- **The use of suites of indicators when measuring research ‘quality’:** While indicators may provide useful indications of impact, there is no direct correspondence between this and research quality – a complex and multi-faceted concept. A key issue relates to the concepts of ‘citation impact’, ‘intellectual influence’ and broader notions including ‘quality’ and ‘excellence’ do not necessarily coincide.

- **Bias against early career researchers:** Bibliometrics depends heavily on an established track record of publication for analysis to be effective. A key problem with this approach is that it may discriminate against early career researchers who do not have a substantial body of publications to their name – unless adjusted measures are used.

- **Gathering accurate publications information is time consuming:** From a practical perspective, there are challenges with collecting accurate publications information. While a number of databases collect basic information on scientific pub-
lications (in the medical research field, Medline, PubMed and the Web of Science are notable examples), a large number of the articles included are not available in full text format. This means that gathering key publication data, including details such as author addresses and funding acknowledgements, can be difficult and time consuming. Second, there is no standard way for researchers to acknowledge their funders, resulting in inconsistencies between journals and research fields. Finally, for papers with a large number of authors, there is no clear way of distinguishing relative contributions made by each individual.

- **Database journal coverage is variable**: Although they are improving, the databases currently available to support bibliometric analysis, their coverage and broad trends in the data that they hold are similarly open to challenge.

- **Identifying authors and their affiliations can be difficult**: Particular forms of bibliometric analysis – particularly those that assess the quality of research produced by research institutions and individuals – depend on an ability to determine accurately the identity and institutional affiliation of the author. This presents challenges.

- **Citation windows need to be chosen carefully**: Selecting citation windows for an evaluation study may have an important impact on subsequent findings.

- **Citation behaviour is variable**: Citation analysis is grounded on the premise that tracing the reference practices of scientists provides us with useful information.

- **Fields can be difficult to define, especially for interdisciplinary research**: Differences between research fields exert important influences over the kind of analysis that can be performed.

- **Attribution is an enduring problem in research evaluation**: The efficacy of bibliometrics as a research evaluation methodology depends to a large extent on the accuracy with which scientific publications can be attributed to particular authors, institutions, grants and so forth. Attribution can be problematic. For example, for publications resulting from collaborative research, it can be difficult to identify which elements of the research can be attributed to particular funders and, furthermore, funders are not always acknowledged on publications.

### Example case studies

#### Research assessment

There have been important moves in Australia and the UK towards integrating bibliometrics into wider research assessment exercises. The Australian Research Quality Framework (RQF) was designed to assess quality and relevance. It aimed to integrate metrics-based approaches much more fully into the analysis, to which there were three elements:

- a context statement from each research group, looking at the culture of research in their area, and some of the most significant advances
- an outline of the full ‘body of work’ produced by the research group, listing all publications and other forms of analysis
- an indication of the best outputs produced by the group – as nominated by the researchers themselves.

The RQF used two bibliometric approaches to support the analysis of data collected:

- standard analysis, based on recognised journal publications by the group
- non-standard analysis, assessing the impact of book chapters and other publications of this kind.

The RQF was replaced (before its launch) by a new research quality and evaluation system, the Excellence in Research for Australia (ERA). ERA incorporates a combination of metrics-based research assessment and expert peer review. The emphasis on metrics-based approaches in ERA is significantly greater than under the RQF, and it incorporates a tentative ranking system for journals.

#### Selecting centres of excellence (biomedical research centres)

As part of the National Institute for Health Research (NIHR) suite of funding streams, the English Department of Health supports research in biomedical research centres (BRCs) – centres of excellence for medical and health research across a variety of
subject areas. The centres are based within the most outstanding NHS and university partnerships in this country. A bibliometric analysis was used to identify outstanding organisations based on highly cited papers in health research fields (as identified through addresses on publications). Further, bibliometric techniques were also used to identify the number of highly cited papers produced by organisations over the period 1995–2001. This provided a visualisation of the continuing success of organisations in health research, and helped to identify which organisations are improving the quality of their research. To identify which organisations would be likely to provide a strong consortium of academic and NHS partners, a network analysis of collaborations between different organisations was carried out using bibliometrics. Once prospective centres had applied for BRC status, the NIHR conducted a series of site visits that were supported by citation analysis of publications submitted by the applicants. A range of bibliometric scores including those on impact levels of research were used. An international panel of experts made BRC selection decisions based on a number of criteria, of which citation analysis was one.

**Summary**

Bibliometric analysis uses data on numbers and authors of scientific publications and on articles and the citations therein (and in patents) to measure the ‘output’ of individuals, research teams, institutions and countries, to identify national and international networks, and to map the development of new (multi-disciplinary) fields of science and technology. In addition, a bibliometric approach generates significant cost savings and can significantly reduce the administrative burden. However, the validity of judgements based on bibliometric information is a key issue and caveats should be considered.
Introduction

The objective of a case study is to explore, explain, or describe an activity. In many respects case studies are self-contained narratives that can be used to illustrate effective practice. In other contexts they can be diagnostic, inasmuch as they describe what works and what does not work. The topics of case studies can also be varied and in this respect they are almost completely flexible. Case studies are a qualitative, descriptive research technique, and provide detailed information about a limited topic or context, with the aim of producing a rich and detailed understanding of that particular area, rather than widely generalisable conclusions. However, groups of case studies together can say more about the wider population if they are carefully selected.

When should case studies be used?

One of the key limitations of case studies is that they are very specific to the context in which they take place. This means it can be difficult to generalise any findings generated through the research as the findings may also be specific to this context. This issue can be addressed somewhat by careful selection of case studies, and/or a sufficiently large case study set. However, caution should always be used when trying to generalise case study findings more widely.

Another limitation of case studies is that they have an element which is inherently subjective. Although they can draw on factual data, often much of the most interesting output of a case study relies on personal interpretation and inferences, and it is difficult to test the validity of findings.

Finally, case studies require a high relative investment per subject or topic. Looking in depth means that it is generally necessary to select a small subset of the total population and to be prepared to ‘overview’ techniques. Accessing this detailed information is relatively costly for individual subjects.

The key advantage of the case study methodology is the ability to access detail and context. When used correctly, case studies can provide rich and deep contextual information, building a full understanding of a particular situation.

Case studies are also able to deal with heterogeneous and non-routine behaviour or circumstances better than many other techniques. As it is an exploratory rather than a prescriptive approach, it is flexible, so it can be creative and innovative, and support the analysis of all kinds of subjects. This can be useful when evaluating research as it allows all kinds of outputs and outcomes of research to be explored, including the unexpected and ‘serendipitous’ outcomes. Where indicators or questions are predetermined beforehand, as in other research approaches, it is possible that some elements may be overlooked.

Case studies are best applied at low levels of aggregation, although this depends significantly on the nature of the case study. The approach is so flexible and wide ranging that it is hard to specify its application. In the research evaluation context, case studies are most likely to focus on a particular researcher or grant, or perhaps a small group of researchers or a stream of work. To access the level of detail required, it is best if this group is small. However, case studies could be applied to a research funding stream or a university research strategy depending on the way in which the case study approach was framed. Equally,
case studies could be conducted at an even higher level, looking at a particular funder for example, but the approach and the outcomes would be significantly different.

As outlined previously, the relative cost per subject of case studies is high. This is because of the level of detail gathered in each case. However, case studies are generally only used to provide examples within a context in a few cases, and are more often used in conjunction with another tool, which can provide the wider picture at a lower cost, such as bibliometrics, or a survey.

Overall, as described by Yin (2009), case studies have a distinct advantage over other methods when a ‘how’ or ‘why’ question is asked. Their strength is providing context and detail to support facts and generalisations produced through other means.

**How are case studies used?**

Given their flexibility, case studies can be used in a wide variety of ways in many different contexts, so it is difficult to identify specific modes of operation that can be succinctly described. A range of classifications are described in the literature (Stake, 1995; Yin, 1993), with some examples as follows:

- **illustrative or explanatory case studies**: for causal investigations; descriptive case studies, which are used to illustrate a particular point, and familiarise readers with an unfamiliar topic; they need not be representative and it is not necessarily valuable to have a large number of case studies for this purpose
- **exploratory case studies**: sometimes used as a prelude to further, more detailed research; they allow researchers to gather more information before developing their research questions and hypotheses; one potential pitfall is if these preliminary findings are taken as conclusions without further investigation
- **descriptive case studies**: involve starting with a descriptive theory; the subjects are then observed and the information gathered is compared to the pre-existing theory
- **intrinsic case studies**: where the researcher has a personal interest in the case
- **collective case studies**: involve studying a group of individuals
- **instrumental case studies**: when the individual or group allows researchers to understand more than what is initially obvious to observers
- **cumulative case studies**: where existing data from different sites or research are aggregated to produce a new set of case studies; this allows a greater amount of data to be studied without significant extra work and may allow for greater generalisation, but there is a risk of non-comparability between case studies
- **critical instance case studies**: look at one or more specific topics or sites of interest; the aim is not to reach generalisable conclusions, but to understand particular situations of interest in detail.

From a research evaluation point of view, case studies can fit into a number of these categories, and from a practical point of view, the classification of a case study only has a limited impact on the approach taken to the case study research given that case study research is such a broad and exploratory approach.

This breadth of applicability equally makes it hard to define a strict methodology for producing case studies. However, Yin (1993) suggests there are five basic components of a research design:

- a study’s questions
- a study’s propositions (if any)
- a study’s units of analysis
- the logic linking of the data to the propositions.

Where multiple case studies are to be conducted, case study selection could also be added to this list. We can use this to consider a very broad method by which case studies can be produced:

1. Select question and proposition of the study.
2. Determine the unit of analysis of the case study. In research evaluation this is most likely to equate to a researcher, or a research project, where consistent motivations and outcomes can be established.
3. Determine the number of case studies to be conducted and the selection procedure. Where gen-
eralisability is sought, the selection criteria can be very important to ensure that the case studies span the relevant dimensions. For example, in research evaluation, when looking at the biomedical sciences, it could be important to ensure you include case studies spanning basic and clinical research.

4 Determine structure of case study. This might draw on logic models, or other formats depending on the proposition of the study in question. When conducting a number of case studies it can be important to use this thinking to develop a template so that consistent levels of data are collected across cases studies to facilitate comparison and generalisation (where required) at the analysis stage.

5 Gather the data for the case studies. This can draw on a range of approaches from interviews, document review and surveys to bibliometric and economic analysis.

6 Structure the data gathered using the logic model or other structure defined previously (potentially with some revisions) to highlight the relevant findings and develop a narrative that conveys the key messages from the case study.

7 Validate the case study, either by allowing participants or other relevant experts to review it, or by triangulating findings to ensure reliability and reproducibility. This is an important step in case study research. Denzin (1984) identified four types of triangulation:
   - data source triangulation, which is a commonly used approach in which findings are compared from different data sources
   - investigator triangulation, where the same case study is examined by different researchers, but this is potentially costly
   - theory triangulation, in which different theories or models are used to interpret results – the extent to which this is relevant depends on the aims and structure of the case study
   - methodological triangulation, when different methods are used to access information, an approach which is also very commonly used in case studies.

8 Analyse the findings from your case studies. This is likely to include cross-comparison between case studies, which may take a variety of forms, in order to isolate themes and messages and context dependent contrasts. Where desirable and feasible, this step may also include drawing out generalisable messages across case studies.

In research evaluation, case studies are not usually used to compare across large numbers of researchers or research projects. Rather, they are used to be illustrative in some way. This may be to generate a set of examples of successful research to use for advocacy purposes, to illustrate examples of certain types of research or research outcomes, or to investigate research processes. In any of these cases the process would be broadly as outlined above. However, the structure, unit of analysis, and types of data collected differ depending on the purpose of the case studies.

One of the key challenges in cases study research is in any generalisation of findings to a wider population. It can be challenging to demonstrate that the cases studies you have produced are generalisable in this way, and this will rely on careful case study selection and thorough validation of case study outputs. In many cases, generalisation is not appropriate, nor even necessary. For example, if case studies are intended to be examples of research success which are to be used for advocacy purposes, this is not a necessary step.

Example case studies

Retrosight

In research evaluation, RAND Europe, in collaboration with the Health Economic Research Group at Brunel University, has developed a case study methodology, Retrosight, which allows us to trace the progress of research from funding to impact. The approach has been used by RAND Europe and Brunel University in a multinational study evaluating the long-term impacts of cardiovascular R&D (Wooding et al., 2011), and is currently being used to assess the returns from mental health R&D (Grant and Wooding, 2010). We have also applied it outside the health field, including in the evaluation of an Economic and Social Research Council (UK) programme, and in the arts and humanities. The Ret-
rosight approach involves developing case studies of R&D efforts that have (or have not) translated into impacts. The case studies are carefully selected to enhance generalisability. The in-depth case studies are developed through a range of methods including key informant interviews, document and literature reviews and bibliometrics. Once the field work is completed the case studies are systematically compared, to differentiate between those that are high impact and those that are not, and then they are qualitatively coded to identify factors that are associated with high quality impact, with sensitivity to the contexts in which the research took place. These observations are then developed into policy implications and actionable recommendations for research funders.

REF impact case studies
In the Research Evaluation Framework (REF) used to assess university research in the UK, case studies are employed to investigate the wider (societal and economic) non-academic impacts of research conducted (HEFCE et al., 2011a). The rationale is that it is not feasible to expect all research to have a wider impact, rather it is expected that universities place value on these impacts and have the infrastructure and support in place to support such processes when they do occur. Furthermore, it is difficult to establish a consistent set of indicators that are widely applicable for these kinds of research outcomes. Therefore, case studies allow the flexibility necessary to understand the wide variety of impact that can occur, while also allowing research institutions to identify illustrative examples of their approach to impact, to support a wider impact strategy statement. The case studies follow a predetermined structure, which is provided centrally, and although necessarily qualitative, they can include quantitative indicators where these are available.

The unit of analysis is not predefined – rather it is determined on a case-by-case basis as required to illustrate the impact under discussion, and hence might cover several researchers working on one research project, or multiple research projects in a stream of work conducted by one or several researchers. All that is required is that there is some consistent thread running through the case study, which is supported by a clear narrative drawing together the different relevant threads of the case study. These cases studies will be comparatively assessed by subject-expert and research-user peer review panels and rankings awarded, which will contribute significantly to decisions around funding allocation.

NIHR: 60 years of NHS research
Case studies can also be used effectively for advocacy purposes. One recent example of this is the work by the UK NIHR marking the 60-year anniversary of the NHS (NIHR, 2008). A series of case studies was produced highlighting stories of research success funded through the NHS. The case studies were focused on illustrating particular improvements in practice and care, so the unit of analysis is not focused on a particular researcher or piece of work. Rather, the unit of analysis is designated as all the relevant research funded by the NIHR and NHS which contributed to the change in practice.
The case studies are selected to illustrate important developments spanning the 60 years of the NHS, two in each decade except for the final ten years, which only include one example. The time frame is demarked by when the research took place rather than when it had an impact. The resulting report has been used for advocacy purposes, to highlight the impact of NHS funded research and to make the case for continued support and funding.

Summary
Case studies can be used for research evaluation in a variety of ways spanning advocacy, analysis, accountability and allocation. They are flexible enough to capture a wide variety of impact, including the unexpected, and are useful in providing the full context around a particular piece of research, researcher or research impact. The primary limitation of case studies is the generalisability of findings, so these are best used where examples are needed rather than for full coverage of a population (unless that population is small). Case studies are often used to provide the contextual information alongside another approach that is less expensive per unit analysed, which can provide generalisable information across a population.
Data mining for research evaluation

Introduction

As organisations accumulate increasing amounts of data in computer-based formats and databases, it has become a challenge to find out how meaningful information can be drawn out from these data. There is increasingly more data stored in databases, but we do not necessarily have a clear understanding of them. Data mining is a research tool to help fill this gap (Fayyad, 2004).

Generally, data mining is understood to be a mechanised process of extracting data from existing databases in a way that allows for the extraction of useful information from the data. It uses algorithms to find correlations or patterns within large amounts of data in related databases. The aim is to present existing data in a meaningful format, which summarises data and reduces their complexity while losing as little information as possible (Fayyad, Piatetsky-Shapiro and Smyth, 1996; Fayyad, Wierse and Grinstein, 2002; and UCLA Anderson School of Management, n.d.). Web of Knowledge (wok.mimas.ac.uk/) is an example of a data mining portal to support research.

As will be discussed, while the investment to develop effective data mining processes might be complex and time consuming, such tools have the potential to reduce the eventual burden of data collection for research evaluation on researchers by making use of information already being collected in databases. Thus data mining could present an opportunity to take advantage of the ever increasing amount of data stored in existing databases for research evaluation to help reduce the burden placed on researchers.

When should data mining be used?

Data mining is used to elicit meaningful information from large amounts of already existing computer-based data (facts, numbers or text that can be processed by a computer (UCLA Anderson School of Management, n.d.). Such analysis of collected data is not new; initially, collected data were analysed through mathematical descriptions (e.g. of data moments such as means, modes, variances); following this, visualisation techniques were used to analyse data sets. The enhanced and growing capacity of computers to store data necessitated the development of robust algorithms to extract information from the databases (Fayyad, Wierse and Grinstein, 2002).

Data mining can be used for prediction and description of information from existing datasets (Fayyad, Wierse and Grinstein, 2002). For prediction, data mining algorithms select data relevant to predicting one variable based on another. For description, data mining is used to find human interpretable patterns describing the data. Several techniques can be used for prediction and description of data: clustering or segmenting data to find groups, separating data into subsets by which to visualise, understand or summarise information; and identifying relationships or patterns between data subsets, or in data sequences and series. The scope of data mining analysis depends on the comprehensive-

---

26 Data mining is about searching for meaning in existing datasets; it must be distinguished from data matching, which is concerned with combining datasets, and can be a step in the process of extracting meaning from existing data.
ness of the available data. Also, algorithms must be made sufficiently robust to cope with imperfect data and extract regularities that are inexact but useful (Witten and Frank, 2005).

In research evaluation, data mining is often used to extract and describe information on research expenditures and outputs from datasets, by drawing on data stored in related databases (Boyack and Jordan, 2011).

**Limitations and challenges**

There are a range of challenges and limitations with data mining, linked to processes for extracting, analysing and presenting information from datasets.

A key challenge in data mining stems from the nature of the datasets to be analysed. The size and dimensionality (number of fields) in databases can increase the search space that must be analysed, requiring more complex algorithms to extract and analyse data accurately. Different techniques can be employed to improve searching of datasets with large scope, including sampling, approximating or identifying irrelevant variables; implementing these techniques requires individuals to understand the content of the data and the algorithm used.

While data mining can prove to be difficult when datasets are large and complex, data mining processes are also challenged when they deal with datasets with incomplete or missing data. This often occurs when datasets were not originally developed with the intention of data mining.

Errors can also arise in the analysis of data through algorithm models. Algorithms can result in overfitting, when they model both the general patterns in a dataset as well as the noise. Assessment of statistical significance of information from datasets is also not necessarily straightforward, particularly when the system passes through many potential models. Finally, often data in datasets are not static and changes over time; these changes could invalidate identified patterns. Strategies to mitigate these potential errors or inaccuracies in analysis include investment in cross validation of data, more complex models, and strategies to conduct additional tests on the data or to identify hidden or missing data.

Finally, the presentation of information extracted and analysed through data mining must be presented in a meaningful and accurate way for human interpretation. Effective visualisation of data from data mining is an ongoing field of inquiry (Fayyad, Wierse and Grinstein, 2002).

**Strengths**

Despite these challenges, data mining can substantially contribute to research analysis. Importantly, data mining provides a way to make sense of large amounts of existing data, by analysing complex and unorganised data, and extracting information from data that were originally outside human comprehension. Automated and computerised analysis of data in data mining can result in reliable and accurate analyses of data, thus potentially improving the quality and strength of research findings.

Another strength of data mining lies in its potential to reduce the burden of data collection required by informants. Data mining tends to shift the burden of research evaluation from suppliers of information to collectors of existing data. By using existing datasets, it minimises the burden of added data collection. However, although data mining can require considerable initial investment to clean datasets, and set up and test algorithms for data extraction and analysis, data mining processes could reduce burden in the long term, through the creation of automated, replicable and repeated processes for extracting and analysing data from datasets. Therefore investment at the outset might result in reduced time and effort required for future analyses.

**Level of aggregation**

The level of aggregation of information derived through data mining depends on the specificity and completeness of data stored in the datasets. Data mining tools could be used with either aggregated or disaggregated data. In data mining, ‘aggregation is the process of computing statistical measures such as means and variance that summarize the incoming stream’ (Gaber, Zaslavsky and Krishnaswamy, 2005).

So far, in research evaluation data mining has been used most effectively to extract aggregated data on research funding and outputs (McAllister and Wagner, 1981; Bourke and Butler, 1999; Butler, 2001). Other studies using aggregated data include...
Jimenez-Contreras, Moya-Anegón and Lopez (2003), Kwan et al. (2007) and Börner et al. (2010). A few studies have attempted to use data mining to extract and analyse data on individual grants and publications (Lyubarova, Itagaki and Itagaki, 2009; Zhao, 2010). In the US, Boyack and colleagues used data mining tools to link individual grants and article data from the US National Institute on Aging (Boyack and Börner, 2003; Boyack, 2004; and Boyack, Börner and Klavans, 2009). In the UK, while Lewison and colleagues analysed data from the science citations indexes and the UK Research Outputs Database (Lewison, 1998; Lewison and Devey, 1999; and Lewison, Grant and Jansen, 2001), grant-article-level analysis was not reported. Acknowledgement data held in National Institutes of Health (NIH) in the US and the UK Research Outputs Database provide acknowledgement data that could be used for data mining. However, Butler (1999) found in a study of funding acknowledgement that, while on the whole acknowledgement data reflected total research output of a funding body, she could not track research to the grant level.

STAR METRICS, a research evaluation framework being piloted and introduced in the US (see Appendix C), aims to use data mining processes to assess research outputs and impacts at the level of individual researchers (Bertuzzi, 2010). Currently STAR METRICS only examines funding and employment data.

Cost or burden
The nature of the approach is that most of the effort required is in setting up the data collection approaches, but that then data collection is relatively automated. The cost or burden of data mining is closely related to the complexity, scope and quality of the database, and the complexity of the algorithm required to extract the information sought. The cost and burden required also vary over time. Investment requirements are greatest at the onset, to identify and clean the data, and to set up the automated data mining system; they should fall dramatically once set up and testing of the automated system is complete. Greater initial investment to develop and test the data mining system can also help to enhance the efficiency and accuracy of the data mining process, and contribute to reducing potential requirements for future improvements or corrections to the process.

It is difficult to estimate the cost of data mining, because of the variation depending on the nature of the data, their complexity, scope and completeness. Experience in the piloting of STAR METRICS in the US in 2009 provides some indication of the relative investment required, compared with other tools for research evaluation. Initial budgeted investment in Star Metrics included $1 million by the National Science Foundation (NSF) and NIH in the first year of operation (NSF, 2010) (compared with their combined annual budget of around $38 billion). Several further agencies have become involved; each will continue to contribute $500,000 a year, in a combination of direct investment and provision of resources, including staff.

Estimates of time investments to set up systems to pull and analyse data for the first level of analysis for STAR METRICS vary, from 20 to 640 hours, with a median of 100 hours. As expected, time investment estimates have fallen over time. Subsequent participants in STAR METRICS are estimated to have to commit between 30 and 100 hours, with a median of 45 hours. Time required each quarter to upload data should be minimal (ranging from 0 to 10 hours, with a median of 2.5 hours). However, these estimates only consider cost and time required to set up and maintain data mining processes for the first level of analysis for STAR METRICS; expanding the system to pull together a wider range of data will increase investment requirements; estimates of the cost and burden of this is not yet known. At present STAR METRICS brings together data from institutions’ administrative systems housing data (e.g. on human resources, indirect costs, vendors and sub-awards).

Data mining as part of the research evaluation toolkit
Data mining can be a reliable and comprehensive way to gather information on the outputs and impacts of research. However, the quality and scope of its use very much depends on the nature and comprehensiveness of the information being gathered in existing databases.
At present, data mining for research evaluation tends to be used with bibliometric methods, to analyse citation databases of research publications, and link publications and their academic uptake, to research funding. The first group of research evaluation studies using data mining tools extracts aggregated data from existing databases for analysis of relationships between research funding and research outputs. Other studies using aggregated data include Jimenez-Contreras, Moya-Anegón and Lopez (2003); Kwan et al. (2007); and Börner et al. (2010). In the 1980s, McAllister and colleagues completed studies on R&D expenditures (measured by aggregated funding amounts), and publication and citation outputs for American colleges, universities and medical schools (McAllister and Wagner, 1981; McAllister and Narin, 1983). In Australia, Bourke and Butler (1999) used data mining techniques to analyse related data on modes of research funding in biological sciences aggregated at sector level, researcher status and academic impact of research. Butler’s work in 2001 used data mining to analyse potential relationships acknowledgement of funding sources on publications and the total research output of a funding body. She found he could not track research back to the grant level through funding acknowledgements, but that acknowledgement data reflected the total research output of the funding body (Butler, 2001). This study by Butler suggests a limitation of data mining in research evaluation: she was unable to analyse research activity at the specific level of an individual grant because of gaps in the stored data.

The second cluster of studies using data mining in research evaluation look at the level of the individual research grant; in these instances data mining again seems mainly to be used for bibliometric studies, linking grant data to individual articles (Lyubarova, Itagaki and Itagaki, 2009; Zhao, 2010). The NIH in the US and the UK Research Outputs Database provide acknowledgement data that can be used to extract data for bibliometrics study. NIH facilitated processes for recording acknowledgement data approximately a decade ago, when the decision was taken to make reference to NIH grant numbers on paper, patent and other publication outputs a grant condition. Drawing on these data, Lewison and colleagues have combined citation data from the Science Citations Indexes with data from the UK Research Outputs Database to extract and analyse national-level impacts in biomedical fields (Lewison, 1998; Lewison and Devey, 1999; and Lewison, Grant and Jansen, 2001); however, they do not report grant-article level analysis. In another instance in the US, Boyack and colleagues published a series of results linking grants to individual articles with data from the US National Institute on Aging, providing evidence of increased citation impact with increased grant size (Boyack and Börner, 2003; Boyack, 2004; and Boyack, Börner and Klavans, 2009).

The scope for applying data mining techniques is limited by the nature of data held within existing databases. Particular types of indicators are likely to be easier to store within datasets. For example, data mining can be useful in gathering existing quantitative indicators of research inputs, outputs and impacts, but it is more difficult to hold and represent qualitative information on processes and the wider context accurately. This suggests the importance of using data mining with a combination of different research approaches for research evaluation. Also, data mining has not yet been tested or used within the context of research evaluation to extract and analyse data on less obvious indicators of impact, for example social or cultural benefits. STAR METRICS anticipates expanding its data mining tool to include social outcomes; its effectiveness in doing so will prove interesting for understanding the scope and limits of data mining for research evaluation.

Therefore, data mining is a useful method to make sense of a number of quantitative indicators, and of vast data, but could usefully be paired with tools that can provide more qualitative methods, such as surveys and case studies, to arrive at a more complete and nuanced understanding of how research contributes to particular outputs and impacts.

**How is data mining used?**

Data mining is part of a broader process of knowledge discovery from databases. Effectively implementing data mining for the purpose of knowledge discovery requires particular steps to be taken to prepare the data, conduct and test the data mining pro-
cesses, and interpret the information. These include (Rokach and Maimon, 2008; and UCLA Anderson School of Management, n.d.):

- developing an understanding of the aims of the research and its application, and any relevant prior knowledge
- selecting datasets for extraction and analysis of data
- pre-processing of the data, including making decisions on what data to use (e.g. aggregated data or a subset of the dataset), data cleansing and data transformation
- choosing appropriating data mining tasks (e.g. classification, regression, clustering summarisation)
- choosing data mining algorithms, the method for searching patterns
- employing the data mining algorithm
- evaluating and interpreting the mined data
- deployment: presenting, documenting and using the knowledge.

Key considerations must be made through this process, to ensure the information analysed accurately reflects the original data, and can be correctly interpreted by the audience. In particular, decisions around how to pre-process the data, and what data mining tasks and algorithms to employ, are critical. In the following sub-sections we discuss the main techniques available for selecting and analysing data through data mining algorithms, some of their key strengths and limitations, and when they can be used effectively.

**Techniques used in pre-processing the data**

The first main consideration when using data mining tools is to decide what datasets to use for the analysis. Preparing data for extraction and analysis of information can require cleaning and transforming data into comparable and consistent formats, as well as ensuring data remain accurate and up to date, accounting for any changes over time. Critical decisions must be made about the scope of the data that should be used: will the algorithm be run on a whole dataset, or will subsets of data be analysed? These are key considerations, particularly with large amounts of data, data distributed over multiple datasets, or with high dimensionality. Also, data mining techniques are increasingly being applied to data streams, data that are multiple, continuous and rapidly changing. In this section we outline methods that can be used to select and prepare data from data sets and data streams, for analysis through data mining (Delen, Walker and Kadam, 2005).

Various techniques have been developed, which can be used with data mining algorithms to select data for the analysis.

Sampling, load shedding, sketching and synopsis data structures involve selection of a subset of data for analysis. Sampling is a process by which researchers use the probability of a data item to be processed to select a subset of a wider dataset for the analysis (Gaber, Zaslavsky and Krishnaswamy, 2005). Sampling can be difficult or inappropriate when the size of the dataset is unknown or if there are fluctuating data rates. Sampling rates should also take into account the parameters of the data rate (the rate at which digital data are transferred) and error bounds.

Load shedding is a technique used to help draw meaning from data streams, in particular by randomly or semantically dropping data elements (Gaber, Krishnaswamy and Zaslavsky, 2004). However, this can carry risks for the analysis, as the relevance of the dropped data to the questions of interest are not necessarily known. When an approximation of the data structure is designed, rather than an exact representation, it is called sketching (Babcock et al., 2002). This can be of particular use for data streams, where an exact representation is not possible; however, caution is required as sketching can compromise the accuracy of data mining. Another method of using approximation to select a subset of data for analysis is to use synopsis data structures. Techniques are used to summarise data, for example in histograms, quantiles and frequency movements. This also results in approximation of the data rather than representing all the characteristics of the dataset.

Aggregation is the technique of using statistical measures to summarise data, for example, mean or variance. Aggregation can be inappropriate for datasets with fluctuating data distributions. The level of aggregation can also affect the specificity and accuracy of the analysis.
Example case studies

STAR METRICS

STAR METRICS (Science and Technology for America’s Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science), a framework for research evaluation being developed by a collaboration between the Office of Science and Technology Policy (OSTP), NIH, NSF, Department of Energy (DOE) and Environmental Protection Agency (EPA) to assess performance on investment of government funded science research in the US is founded on a data mining approach. Data mining was introduced with the aim to minimise burden on researchers and create a comparable and reproducible set of data on performance.

STAR METRICS is the first attempt to use data mining to relate data on research funding and publication, with data on economic and social impacts (e.g. job creation, economic growth, workforce outcomes, scientific knowledge, social outcomes) for research evaluation. It is being implemented through a two-step approach, following a pilot conducted in 2009. At Level 1, the framework includes measures of job creation by looking at who is employed through the funding provided. At Level 2, STAR METRICS is likely to be expanded to develop wider indicators covering the impacts of research on economic growth, workforce outcomes, scientific knowledge and social outcomes. The exact metrics that will be used to measure impact in these different areas are still being developed.

At Level 1, data are collected on the number of scientists, including graduate and undergraduate students, as well as researchers, supported by federal science funding. This includes job creation through sub-awards, sub-contracts and overheads. The data collected at Level 1 fall into the following categories: information on awards; information on individuals; information on indirect costs; payments to vendors; and sub-contracts and sub-awards. At Level 1, a key principle of the approach is to use the existing administrative systems of each institution; STAR METRICS works with the institutions to develop an automated system which extracts 14 key data elements from the data they already collect.

Data mining tasks and algorithms: techniques for extracting and analysing data

Different data mining methods are available to search, assess and describe datasets. Appropriate selection and development of techniques for data mining requires understanding and testing the techniques and the content of the data.

These are some of the most common data mining techniques:

- **Decision trees**: Decision trees are predictive models that can represent classifiers and regression models. They use univariate splits to compare a particular attribute value with a constant. This results in a simple model which is easy to comprehend.

- **Classification rules**: Data are mapped out according to several predefined classes.

- **Regression**: Data are mapped to a real-valued prediction variable.

- **Clustering**: A finite set of categories or clusters is used to describe the data.

- **Summarisation**: A compact description is developed for a subset of data.

- **Dependency modelling**: A model is developed to describe significant dependencies between variables.

- **Change and deviation detection**: Previously measured or normative values are used to identify significant changes in data.

Algorithms are then constructed to apply these different methods to the datasets. Algorithms for data mining include three primary components: model representation, model evaluation and search. Model representation is the language to describe discoverable patterns; model-evaluation criteria are quantitative statements that identify the extent to which a particular pattern (model and its parameters) aligns with the goals of the process (e.g. criteria could include predictive accuracy, novelty, utility, understandability of the fitted model); and the search method includes both the parameter search and the model search, which optimise the model-evaluation criteria given observed data and a fixed model representation.
Significantly greater and more complex data collection is anticipated at Level 2, which has yet to commence. Level 2 proposes to use data mining in combination with natural language processes. It will involve the creation of a new online dataset, through which scientists can upload a profile. The profiles would become the basis for pulling information into research funding applications, progress reports and other documentation. Data mining algorithms for Level 2 analyses will be developed through a staged approach. First, the core focus is to develop the web profile system, a core information platform, and to draw together existing datasets and databases. Second, inputs to the profile system and the system’s automation are to be improved, and approaches are to be developed which generate connected data about impacts in a standardised way. Finally, data mining processes are to be further developed to include information such as interactions with the international scientific community.

**Coverage**

STAR METRICS is planned to be applied potentially across all government funded science research organisations in the US; at present, coverage is much lower and participation in the databases to be used for the data mining is voluntary. There are currently around 80 academic institutions at various degrees of participation at Level 1 of STAR METRICS. Also, only around 120 of the total 4,800 universities in the US fall into the group of Tier 1 research universities, which conduct nearly all federally funded research. These are the primary targets for STAR METRICS. It is estimated that with the universities currently committed to the scheme, STAR METRICS covers around 40 per cent of all federally funded university research in the US. Because participation in the databases used for data mining with STAR METRICS is voluntary, it is likely that the databases will result in a biased sample. They are likely to favour particular types of institutions which stand to gain most from participation in this type of programme; this may not align with the primary target universities for STAR METRICS.

**Level of aggregation**

STAR METRICS aims to use individual researchers as the unit of assessment (Bertuzzi, 2010). Questions still remain on how the data will be aggregated, and how the information processed will be used.

**Cost or burden**

As discussed above, the cost and time estimates for fully implementing Level 1 and Level 2 of STAR METRICS is uncertain. Initial investment to set up and pilot Level 1 was $1 million by NSF and NIH in the first year of operation (NSF, 2010) (compared with a combined annual budget of around $38 billion), as well as contributions of $500,000 from the other agencies which have become involved (in a combination of direct investment and provision of resources, including staff). Time investments vary considerably, from 20 to 640 hours, with a median of 100 hours, but are expected to fall over time. Time required each quarter to upload data should be minimal (ranging from 0 to 10 hours, with a median of 2.5 hours).

**Presentation of data**

Visualisation is considered an important part of the process to make the gathered data useful and intuitive. The Level 1 data have been presented in a number of formats. One is map-based, in which jobs supported by federal science funding are mapped across regions, an example of which is shown in the chapter on STAR METRICS in Appendix C.

Stanford Physics Information Retrieval System (SPIRES) is used by the US NIH to relate publication data to grant project data. It only uses data mining techniques to consider impacts on scientific knowledge, as indicated by citation data. SPIRES maps publications to NIH grants, using automated text manipulation methods to extract and reformat grant numbers cited in publications. They are then matched to NIH grant data and the quality of the match is rated by the system (Rockey, 2011). SPIRES provides a tool to analyse publication and project data, by translating National Library of Medicine (NLM) publication data into a relational database that can be directly linked to electronic research administration grant data.
Steps in the application of the tool
Publication records that contain either the full NIH institute abbreviation or a two-letter US Public Health Service (PHS) organisation code are first downloaded in bulk from PubMed. Following this, relevant data (PubMed IDs, grant agency, acronym, number) are extracted and loaded into a relational database. A parsing engine is used to decode the grant ID string to identify the three valid NIH project number components. A challenge at this stage is that grant ID values are presented in different formats in the publications.

Project numbers are then matched to PubMed IDs, and a match case scoring system used to assess match quality, based on two conditions (the components of the NIH project number that could be unambiguously matched, and the number of core project numbers that are either fully or partially matched). This information is presented for interpretation in a table. Finally, Scopus, a citation database, is used to link citation data.

History
Dr Ben Van Houten began the SPIRES project in 2001, at the National Institute of Environmental Health Science (NIEHS), with the aim to match existing data on publications from PubMed with their NIH extramural grant numbers, held in the electronic research administration database. The first version of SPIRES was completed in 2002, with data going back to 1995 from the NIEHS, National Institute of Child Health & Human Development (NICHD), National Institute of Ageing (NIA) and Fogarty International Center (FIC). Data had to be manually downloaded and updated monthly. The second phase of the SPIRES project aimed to expand the database to all NIH institutes and centres, and to develop a reliable automated data mining process that could be updated daily (Jordan, 2007). The scope of SPIRES has grown to 24 NIH institutes and centres, but the system continues to focus on linking grant project data with publication data, linking data on projects, journals and ISI journal impact factor data. Additionally, the daily automated updates account for new publications, and changes to the last revised date and to MeSH terms indexing data for already identified publications.

Use for research evaluation
SPIRES supplies data for the NIH Research Portfolio Online Reporting Tool Expenditures and Results (RePORTER) system, whose website (http://projectreporter.nih.gov/reporter.cfm) attempts to extract and present existing data for systematic research evaluation purposes, available since mid-2009. This allows queries to be made of NIH funding and publication data. The ExPORTER extension makes data available for bulk download. Available data date back to 1985 and include tables for projects, links from articles to projects, articles and patents.

Limitations and suggestions
These are some of the challenges in the use of data mining in SPIRES:

- **Generation of false positive matches between grants and article:** The fraction of false positives is unknown but expected to be small, as the system excludes matches to unawarded grants. It also excludes approximately 1 per cent of the data in which articles appeared to have been published before the grant was awarded.

- **Incomplete data in the datasets:** PubMed does not index all articles that acknowledge NIH grants; article counts and publication rates should be considered lower bounds. Similar incomplete data could occur in the citation database also.

- **Inconsistent records for grant acknowledgement:** For example, most records in PubMed do not include suffix information and cannot be linked to individual grant years.

Summary
Data mining techniques present an exciting opportunity to ground research evaluation in existing data held within a growing number and size of databases. Setting up processes for effective data mining transfers the burden of research evaluation onto ‘collectors’ of data, with most effort required in setting up appropriate algorithms for extracting meaningful information from large amounts of data, held in multiple datasets. These processes open up opportunities to discover hidden knowledge, and use this to inform expected knowledge.
More widely, the possibilities and applications for data mining are growing: new tools for data mining include graphical user interfaces, video mining (e.g. work to blend speech recognition with image and natural language processing, see work at Carnegie Mellon University), web usage mining and so on. However, as data mining becomes more widely applied and the data that can be included in the processes expand, technical and social concerns are materialising, for example, technical concerns about relative strengths and weaknesses of different tools, or social concerns about privacy of individuals. Also, using data mining for research evaluation effectively will require data to be recorded in databases (e.g. acknowledgement data) and for processes to be developed to combine data held in different formats. Such challenges are likely to continue to be of concern in the application of data mining in the future (Glover et al., 2010).
Introduction

Document review is a methodology that is useful for gaining a broad overview of the material produced on a particular issue and to identify ‘what is known’. A document review should be tightly focused on the key questions to be addressed to avoid ‘trawling for evidence’, which is time consuming and wasteful.

When should document review be used?

Document review is often the first step in gathering research evidence. The aim is to evaluate and interpret all available research that is relevant to a particular question. The challenge for any such review is to identify and access the literature sources that address the specific issues of interest, avoiding the potentially vast amount of less relevant information. The information that can be obtained is necessarily limited to only that which has been published and is available to the researcher. Some of the most interesting information may not be publicly available, so accessing these data can be a challenging element of the process. The advantage of document review is that it puts very little burden on research subjects, and is relatively low cost provided excessive trawling of irrelevant data is avoided. Often this is a necessary step before employing other research tools, for example to establish what the key issues are to be addressed in a survey and hence which questions need to be asked, or in identifying the key stakeholders before interviews can be conducted.

How is document review used?

Document reviews typically involve two steps – first analysing the existing evidence base, which includes identifying what literature exists, and assessing the content and suitability of the literature, and identifying the gaps. The second stage is to develop an interpretation so the emergent meanings can be articulated, their relevance clarified, and the findings refined, enhanced and strengthened. Through this synthesis process such reviews can add to the existing pool of knowledge, although no new primary data are collected.

In research evaluation, documents that are likely to be relevant for review include research strategy and policy documents, research proposals, previous reviews of the research, the key publications resulting from the research, end-of-grant reports, annual reports for researchers or grants, and wider commentary on the research in reviews of the topic, the media or elsewhere.
Introduction

Economic evaluation is a comparative analysis that examines the costs (inputs) and consequences (outputs) of two or more policies, actions or interventions. It adds to a descriptive case study a quantification of economic effects, such as through benefit-cost analysis, to estimate whether, and by how much, the benefits of a project exceeding its costs providing useful information for decisionmaking. Economic evaluation studies therefore provide a structured and systematic way of helping decision-makers choose between competing or alternative ways of using finite resources.

There are two distinguishing features of economic evaluation studies. They:

• deal with comparing costs and consequences
• are concerned with choices.

The latter distinguishes economic evaluations from economics. Whereas the discipline of economics tries to explain the choices and behaviours of individuals and organisations, economic evaluation studies seek to inform choices that must be made by policymakers or other decisionmakers.

The purpose of economic evaluation studies is twofold. They assess whether:

• the benefits from the policies under consideration are greater than the opportunity cost of those policies (compared with alternative uses of the resources)
• allocative and technical efficiency is achieved.

When should economic analysis be used?

There are several different types of economic evaluation. The three most commonly used are cost-benefit analysis, cost-effectiveness analysis, and cost-utility analysis. They are defined as ‘full’ economic evaluation studies. The main distinguishing factor in these three types of economic evaluations is in the way the consequences of respective policies are expressed. In contrast to cost-benefit analysis, cost-utility analysis and cost-effectiveness analysis, implicitly assume that one of the programme or policy alternatives will be undertaken regardless of its net benefit. Hence, cost-effectiveness analysis may lead to a decision to undertake a programme, intervention or policy that does not pay for itself because the technique assumes that the output (measured in health effects) is worth having and the only question is to determine the most cost-effective way to achieve it (Drummond et al., 2005).

Cost-benefit analysis is best used when the goal is to identify whether the benefits of a programme or policy exceed its costs in monetary value. Since cost-benefit analysis converts all costs and benefits to money, the advantage of cost-benefit analysis over cost-effectiveness analysis or cost-utility analysis lies in the ability to make decisions about a policy or programme in stages (rather than comparing two alternatives simultaneously) and with or without the constraints of a fixed budget.

Cost-effectiveness analyses are used when costs are related to a single, common effect that may differ in magnitude between the alternatives. For example, if a policy interest concerns the prolongation of life after renal failure and the interest in compar-
Appendix D

145

ing the costs and consequences of hospital dialysis with kidney transplantation, the outcome of interest is common to both programmes: life years gained.

Cost-utility analysis is most appropriate when costs are related to alternative policies that have multiple dimensions and outcomes, and where quality of life is either the most important outcome or an important outcome among others (e.g., survival).

There are a number of limitations and caveats to using economic evaluation.

Cost-effectiveness analyses and cost-utility analyses should not be used if:

- data on all alternatives are incomplete and/or non-comparable
- there is no formal periodic budget allocation process during which all alternatives can be assessed simultaneously
- there is a need to know whether a particular goal of a programme or policy is worth achieving given the social opportunity costs of all the resources consumed in its implementation (assuming the social costs are also known)
- there is a need to capture effects that spill over to other persons (positive or negative), known as externalities in economics (e.g., health effects of air pollution such as chronic obstructive pulmonary disease, or asthma).

Cost-benefit analysis should not be used if:

- there is a need to know only the price of achieving a particular goal or outcome, whether it is the incremental cost of a life year saved, a case of disease detected, or a quality-adjusted life year (QALY) gained
- decisions on allocative efficiency are not required, rather it is assumed that a particular policy or programme will be implemented
- the client focus of the expected outcome is narrow
- assigning monetary values to outcomes is neither appropriate nor possible.

How is economic analysis used?

Cost-benefit analysis

Cost-benefit analysis expresses assessed impacts in monetary units and is often the most useful type of economic analysis for decisionmakers. However, it is also the most difficult type of economic evaluation to conduct. This is because it compares the expected costs and benefits of two (or more) alternative policies, actions or interventions where all items are expressed and valued in monetary terms. The difficulty lies in the fact that measuring costs or benefits and valuing them in a currency requires many different skills and associated professionals. When assessing the spectrum of research from upstream through to downstream research, cost-benefit analysis use is more feasible in evaluating applied research and technology development programmes than basic science programmes where the ultimate outcomes and impacts may be decades away and difficult or impossible to capture. However, even with applied research and technology development projects, there may be difficulties in estimation related to a project’s distance from the market. Generally, the further upstream of the market a programme is positioned, the more complicated becomes the task of apportioning costs and disentangling the contributions of various contributors to the development of the eventual technology, and of estimating downstream economic benefits.

Cost-effectiveness analysis

Cost-effectiveness analysis compares the outcomes and results between alternative policies, actions and interventions that affect the same outcome. Thus, cost-effectiveness analysis estimates expected costs and outcomes of policies. It expresses the outcomes in a single dimension measure (natural effectiveness units such as life years gained or saved, cases prevented, disability-days saved, and so on). The outcomes in cost-effectiveness analysis could be intermediate or final, but they are nevertheless single, policy- or programme-specific and unvalued. In the case of health, for example, intermediate outcomes may be symptoms or risky behaviours, whereas final outcomes may be cases or deaths. Ultimately, this method produces a summary measure, a cost-effectiveness ratio, for a particular policy, action or intervention in the form of cost per outcome achieved (e.g., cost per cases prevented, cost per deaths avoided, cost per quitter, cost per abstinent, and so on). A cost-effectiveness analysis is used primarily to identify the strategy, or policy, under a fixed budget,
Cost-utility analysis

Cost-utility analysis compares costs and benefits of alternative policies (or interventions and so on) to help decisionmakers determine the worth of a policy or programme by reference to an external standard (usually a fixed budget). As in the case of cost-effectiveness analysis, cost-utility analysis relates to constrained maximisation. However, cost-utility analysis differs from cost-effectiveness analysis with regard to outcomes. In cost-utility analysis the investigation may be single or multiple, are generic (as opposed to policy- or programme-specific) and incorporate the notion of value. Hence, cost-utility analysis is more useful than cost-effectiveness analysis to decision-makers with a broad mandate because cost-utility analysis has broad applicability. Cost-utility analysis is viewed as a particularly useful technique because it allows for quality of life adjustments to a given set of outcomes, while concomitantly providing a generic outcome measure for comparison of costs and outcomes in the alternatives examined. Cost-utility analysis produces an integrated single measure, the QALY, that accommodates the variation in preferences individuals or society may have for a particular set of outcomes by capturing gains from reduced morbidity (quality gains) and reduced mortality (quantity gains). The result of cost-utility analysis is typically expressed as the cost per QALY gained by undertaking one policy or programme over another. The different types of economic evaluations provide varying forms of analyses. These are some of the standard steps of a cost-benefit model that are recognised as best practice (Canadian Treasury Board, 1976):

- **Examine needs, consider constraints, and formulate objectives and targets**: It is important that all economic evaluations clearly indicate the perspective from which costs and benefits will be assessed.
- **Define options in a way that enables the analyst to compare them fairly**: When an option is assessed against a baseline case, it is important to ensure that the baseline case has been optimised.
- **Analyze incremental effects and gather data about costs and benefits**: It is helpful to specify all the costs and benefits over time in a spreadsheet.
- **Express the cost and benefit data in a valid standard monetary unit of measurement**: This step involves converting nominal dollars, pounds, Euros and so on to a constant currency, so the cost-benefit analysis uses accurate, undistorted prices.
- **Run the deterministic model**: The single-value costs and benefits are taken to be certain.
- **Conduct a sensitivity analysis to determine which variables appear to have the most influence on the net present value**: This step involves considering whether better information about the values of these variables could be obtained to limit the uncertainty, or whether action can limit the uncertainty (eg negotiating a labour rate).
- **Analyze risk**: This refers to risk that arises from uncertainty in the data. The analysis is made by considering what is known about the ranges and probabilities of the cost and benefit values and by simulating expected outcomes of the investment.
- **Identify the option which gives the desirable distribution of income**: This is done along a chosen dimension such as income, class, gender, region and so on – whatever categorisation is deemed to be appropriate.
- **Analyze factors that cannot be monetised**: Considering all of the quantitative analysis, as well as the qualitative analysis of factors that cannot be expressed in dollars, make a reasoned recommendation.

**Example case studies**

**Estimating the economic value of medical research in the UK**

Despite a growing international interest in understanding the health and economic returns on investment in medical research, there is relatively little formal analysis, particularly in Europe. An economic evaluation was used to analyse returns from such investment in the UK. The evaluation empirically estimated two main elements of economic returns from medical research – the value of health
Estimating the impact of research investment in virus-based vaccines

A study by assessing the value of health R&D in Australia evaluated specific R&D activities that led to improved quality of life and reduced mortality (Access Economics, 2008). A cost-benefit analysis of Gardasil, a vaccine which protects against certain types human papillomavirus, introduced in 2007, demonstrated significant benefits. The estimated cost of developing Gardasil was A$1.3 billion and it was estimated that the total cost to society of a person contracting cervical cancer was A$1.63 million. The R&D benefit of Australian research was estimated at A$63 million per year compared with a cost of A$8.5 million per year, an estimated cost-benefit ratio of 1:7.5.27

Summary

Economic evaluation takes a number of different forms, depending on the extent of monetisation of costs and benefits to be analysed and/or compared. Each type of economic evaluation has a different purpose and this determines the conditions under which it is used. Given the different economic perspectives that can be taken, there is no single way to conduct an economic evaluation.

27 At the national level the study estimated that 1992/93 and 2004/05, expenditure on Australian R&D returned a net benefit of approximately $29.5 billion. Lateral Economics (2010) found that for every dollar invested in Australian health R&D, an average of $2.17 in health benefits is returned.
Introduction

Logic models are graphic representations of the essential elements of a programme. They encourage systematic thinking about the programme and its underlying assumptions, and to some extent about other external factors affecting the achievement of the ultimate outcomes. They facilitate the identification of linkages between elements of the programme and can be used to identify causality and expose gaps. Logic models can serve as an ideal guide to planning, monitoring and evaluation.

The most basic logic model depicts how a programme works. It describes how inputs or resources feed into a sequence of activities, and how these activities are linked to the expected results of a programme. A logic model illustrates the connection between planned work, which describes the types of resources (or inputs) and the activities that need to happen to carry out a programme, and intended results, which include all the programme’s results over time: outputs, outcomes and impacts (W. K. Kellogg Foundation, 2001). The intended results of a programme are influenced by the programme environment. As programmes operate in open systems, environmental factors can increase the likelihood that a programme will succeed and at the same time impede the success of that programme. Thus, specifying and thinking ahead about these influencing factors is a step forward in developing a logic model.

When should logic models be used?

There are also potential limitations with using logic models:

- Logic models cannot always be applied to programmes (McDavid and Hawthorn, 2006). For example, this could be the case with particularly turbulent programmes. Under such circumstances, developing logic models might not be a useful and effective way of understanding the dynamics of a programme, or how planned work relates to intended outcomes.

- Logic models cannot capture the counterfactual. They do not show what would have happened without the intervention in place, or if another intervention had been implemented.

- Like programmes, logic models are dynamic and time limited. Logic models are an instant picture of a programme at a specific moment in time. In other words, as the programme develops and changes, so too will the logic model. A logic model is a work in progress, a working draft that can be refined as the programme unfolds. If a logic model is not updated, it may become obsolete and, potentially, misleading.

- Logic models are linear and might therefore miss feedback loops and fail to reflect learning across initiatives. To communicate these feedback loops, evaluators may highlight them during interviews or workshops. However, occasionally, it might not be possible to capture or understand feedback loops directly through logic models, since logic models combine goal hierarchy and time sequence.

- Logic models sometimes identify programme ‘reach’ poorly. They treat the ‘who’ and the ‘where’ on a rather secondary level, focusing more on the causal chain between the different elements of the logic model than on reach. Montague (1998) identified some problems when models focus poorly on reach:
logic models, together with the organised approach to collecting and collating information, generally provide a clear picture of what you planned to do and why. This feature of logic models enhances the case for investment in a particular programme.

• Logic models reflect group process and shared understanding. Ideally, logic models should be developed in conjunction with the various stakeholders of a programme. The involvement of stakeholders is not only key to reviewing and refining the programme concepts and plans, but also contributes to getting everybody’s involvement and buy-in.

How are logic models used?

What are the different ways in which this tool can be used and when they are most appropriate?

Figure A21 shows a basic logic model. Although it reads from left to right, developing it follows a retrospective approach. In other words, the evaluator should first start by specifying what will happen (the outcome or impact), and then work backwards to identify the various elements of the logic model. Once the initial logic model has been developed, the evaluator might want to validate and identify potential gaps or weaknesses by following the chain from left to right and testing it step by step.

Logic models are useful tools for framing evaluation questions, programme planning and implementation, and programme evaluation.

Framing evaluation questions

A logic model is a simple but representative tool for understanding the context in which a programme works. By addressing questions that explore issues of programme relationships and capacity, evaluators can better understand how the programme relates to the wider economic, social and political environment of its community. Furthermore, logic models are a helpful tool for identifying potential gaps or issues during implementation that need to be addressed to deliver the programme as planned (programme planning and implementation), and determine the programme’s progress towards desired changes in individuals, organisations, systems and communities (performance evaluation).
Measuring research: a guide to research evaluation frameworks and tools

Programme planning and implementation

One of the most important uses of the logic model is in programme planning and implementation. A logic model illustrates how a programme will work, identifies the factors that potentially will affect the programme, and enables the planner to anticipate the data and resources (inputs and activities) needed to achieve success. It forces the evaluator to clarify its theory of action. At the same time, by providing a good conceptual ‘snapshot’ of the programme, the logic model serves as a useful planning tool for developing an adequate programme strategy. This will include the identification and collection of data for programme monitoring.

Performance evaluation

Performance in the private sector is often measured as financial benefit or increased sales. Traditionally, governments also used to describe programmes in relation to their budgets. However, financial resources spent on a project do not necessarily reflect on the programme’s success or failure. Consequently, governments and non-government organisations have adopted new ways of assessing performance and understanding what progress has been made towards the intended outcomes. A programme logic model can provide output and outcome measures of performance as relevant indicators. It is a useful tool for presenting information and progress towards goals previously set.

Factors to be taken into account before developing a logic model

Before starting to develop a logic model, some important factors need to be taken into consideration:

- Logic models are best used to depict major, recurring items within a programme, rather than individual items. The logic model should provide a macro perspective as well as an overview of the interactions between the different programme elements. As a result, focusing too much attention on the small details of the programme might be distracting and ineffective.
- The size and the level of detail of a logic model can vary, but overall it should be such that read-
Finally, after completing and reviewing the draft logic model with the stakeholders, it should be revised and validated as a workable model of the intended processes and outcomes of the programme. This would be the final logic model. The evaluator must remember that a logic model can be represented in multiple ways (e.g., different levels of detail), so there may not always be a common understanding of how the model should look.

Document reviews, interviews, and focus groups are most commonly used to populate a logic model, though other methods could also be employed. Regardless of the method selected, the development of a logic model always involves a significant amount of professional judgement.

**Specific steps in logic modelling**

To create a logic model, the first step is to reflect on the situation of the programme. As explained earlier, an outline of the situation should provide a good overview of the relevance of the project, providing a statement of the problem and a description of who is affected and which other stakeholders might be interested in the programme.

Once the elements of the programme situation have been identified, it is important to reflect on what is ultimately intended by the programme—the intended outcomes and impacts. Then there is a backward process linking the various elements of the logic model.

To populate the logic model, data need to be collected in advance. To collect such data, the following steps should be considered:

- Review any documents that describe the programme and its objectives. These can include policy documents, working papers, memoranda and so on.
- Meet and interview programme managers and programme stakeholders to learn more about the purposes and activities of the programme, as well as to get further information about how the programme will meet the intended outcomes.
- Construct a draft logic model based on the information collected during the first two steps (e.g., following the structure of Figure A21).
- Present the draft logic model to programme managers and stakeholders (ideally the same people interviewed) as part of an iterative process. It may be necessary for the evaluator to explain what a logic model is and how it clarifies the structure of the programme and its objectives. Once the model has been presented, discuss it with programme managers and stakeholders to help to fill any information gaps and, if necessary, to fine-tune the model.

**Example case studies**

**Measuring the impact of health research**

An evaluation of the 10–15-year paybacks of research grants by the Arthritis Research Campaign (ARC) used a multi-dimensional payback framework with the aim of informing funding and governance decisions by comparing the value for money of different types of grants, and identifying success factors in the translation of research into practice. The system that was developed for evaluating arthritis research allowed ARC to stimulate and manage the exploitation of research advances so that they translate into outcomes of practical benefit to people with arthritis. The research was structured around the Buxton-Hanney Payback Framework, which consists of a logic model (Figure A22) of the research process and a five-category classification system for the benefits of research: knowledge production, capacity building, informing policy and product developments, health benefits, and broader economic benefits. The evaluation used these categories to classify payback (outputs and outcomes of the research funded by the ARC) and determined indicators for research benefits.

The logic model was used in evaluation methodology to understand input–process–output relationships and to break down research programmes into parts. A key element of the model are the interfaces between the research and the stock or reservoir of knowledge and the political, professional and soci-
Figure A22
The Buxton-Hanney Payback Framework

Measuring research: a guide to research evaluation frameworks and tools

The political, professional, and industrial environment and wider society. The degree of permeability at the interfaces is key in determining how far the research conducted meets the needs of potential users and how far the findings are then brought to the attention of users and absorbed by them. Here the roles of formal arrangements and informal dissemination and networks are important. An important feature of the framework, however, is that recognition of the importance of such activities is embedded into the wider perspective. Therefore, rather than trying to follow all the network and dissemination activities, payback assessments can focus on key impacts on policy and practitioners and then trace backwards to identify the importance of the specific networks in achieving that identified impact.

Based on this approach, the evaluators of research funded by the ARC were able to provide insights including that short, focused project grants provided value for money. The resulting significant value for money of these projects illustrated that ARC should continue to include smaller, focused grants as part of its wider research funding portfolio.

The payback framework proved to be effective in capturing the diverse range of ARC’s research outputs and outcomes, and in identifying the cases where the research had been translated to benefit people with arthritis. Given the appropriate management information, there is good evidence, therefore, that the payback framework developed for ARC could be operationalised to monitor the returns from arthritis research prospectively. If applied prospectively, the framework could be used to inform the granting of the recommended translation and partnership awards.

Measuring return on investment in health research
A number of organisations in the Canadian health research landscape are implementing the Canadian Academy of Health Sciences Framework (CAHS)\textsuperscript{28} to assess the impact of the health research they fund and determine how to improve returns to investment in health research.\textsuperscript{29} The CAHS framework

\textsuperscript{28} Details of the CAHS framework are provided in Appendix C.

\textsuperscript{29} The CAHS framework is also in use in countries outside Canada, including the US, the UK and Spain.
is a logic-based model that aims to capture specific impacts in multiple domains, at multiple levels and for a wide range of audiences. It provides a common approach to health research funders in tracking health-research impacts. Alberta Innovates Health Solutions (AIHS), one of the health research funders currently using it, is assessing a number of impacts including the research impact of the grantees that are funded by the organisation. One of the particular strengths identified by AIHS is the clarity the framework brings to the user in how to categorise the multitude of outcomes and impacts that can result from research. AIHS has also found that the logic model with the articulation of the theory of change from research to health impact has significant benefits and facilitates the ability to test contribution analysis.

The set of indicators and metrics proposed for use by the CAHS framework in evaluating the returns on investment in health research provides a useful start for thinking about indicators that focus on an organisation’s specific goal. AIHS has developed new measures for collaboration activity, informed decisionmaking, emphasis on innovation and a narrative for telling the impact story at the level of the project and programme that directly address their mandate.

One of the caveats identified with the framework is that it cannot resolve all complexities easily or immediately. If applied prospectively, it requires time and, depending on the scope, it could become costly to apply it. Another issue relates to producing results that are comparable and enable effective communication of results of the funding to policy and decisionmakers. On the whole (and although still under implementation) AIHS is finding the comprehensive nature and balanced approach of the framework particularly appealing. AIHS has also found the framework to be particularly useful in providing scope for thinking about new ways of capturing impact.

**Summary**

Logic models are graphical representations of the inputs, activities, outputs, outcomes and impacts of programmes or projects. They allow users to think systematically about a programme’s elements and how they link together, identifying potential gaps, developing a common understanding of the programme among stakeholders, and organising information in a practical and structured way. Therefore, logic models are appropriate for framing evaluation questions, programme planning and implementation as well as performance evaluation. Logic models are context specific. If programmes are particularly complex, with significant feedback loops and highly changing dynamics, the evaluator might want to consider using a different approach.
Introduction

In its most basic form, peer review involves external (and sometimes internal) academic reviewers in the process of evaluating research. This can serve a range of purposes, from deciding which applications to a funding body are awarded financial support, to assessing which research is suitable for publication in a journal, or reviewing the performance of a researcher or institution over a period of time.

There are typically two stages to any peer review process:

• a review stage, in which a review is conducted, by individuals or committees, to assess quality; dimensions of quality included in the assessment vary depending on the nature and aims of the review process
• a decision phase, in which a final outcome is reached and the results of the review process are relayed to the participants.

The key overarching principle is to use subject experts working in the field, to provide a review of the work of their peers, the rationale being that they are uniquely qualified to understand the subject and the challenges involved with the work, and hence are best placed to assess the quality of the work of others.

When should peer review be used?

One of the key advantages of peer review is its credibility with the academic community, and more widely. It has been widely used and tested and support for peer review among academics and researchers is strong (Wooding and Grant, 2004; Ware, 2008). This view is reflected in the media, too, with acknowledgement that while peer review may not be ‘perfect’, the alternatives still leave much to be desired (Hawkes, 2008). Perhaps most importantly, confidence in the system among policymakers and research funders is strong (RCUK, 2006). This buy-in is crucial in acceptance and credibility of the decisions resulting from such assessments.

One of the key criticisms levelled at peer review is the cost and burden of the process for applicants and reviewers. However, this burden differs significantly depending on the outcomes of the process. For example, where peer review is used in funding allocation, the cost and burden of the process must be taken in proportion to the funds allocated as a result. In the UK, the 2008 Research Assessment Exercise, which was primarily a peer review process, cost around £60 million to administer, including costs to the participating institutions, but it led to the distribution of at least £6 billion in quality-related research funding,30 so this cost was less than 1 per cent of the total allocation. By comparison, the administration costs of the UK research councils, the publicly funded agencies responsible for coordinating and funding research for projects, are 4 per cent of their total budget (RCUK, 2006), not taking into account the costs to institutions.

In addition to these criticisms about the efficiency of peer review, questions remain about its effectiveness. It is not clear to what extent peer review achieves the aims of the review process. It is not clear that effectiveness has been demonstrated,

---

30 QR funding from HEFCE alone for the year 2012/13 will be around £1.6 billion. See HEFCE (2012).
so although there is no clear evidence that peer review overall is ineffective, equally there is little evidence that it is effective. Part of this is a result of the predominance of peer review in research evaluation, so there are few alternative approaches, which limits the possibility to compare the effectiveness of peer review.

Another potential limitation of peer review is the inherent conservatism of the approach. Research that is more innovative tends to appear less well supported by the previous literature and may be disadvantaged in peer review. Reviewers who are academics are more likely to favour tried and tested approaches that are known in the field (Roy, 1985; Braben, 2004). Similarly, young researchers may be disadvantaged as they are not able to demonstrate a substantial body of previous publications to show their ability. This is supported by an NIH review showing a decrease in success rates for early career researchers, which cannot be accounted for by annual variations in the overall quality of applications (NIH, 2008).

Peer review can be more difficult to apply in cases where the assessment criteria differ from a simple assessment of academic excellence in a well-defined academic field. For example, when research is interdisciplinary, it can be challenging to identify the appropriate ‘peers’ to review research. Similarly, when research studies look at wider outcomes of research, such as societal impact, rather than just academic excellence, it is questionable whether a group of academic peers is best placed to make this assessment. In the UK Research Excellence Framework the proposed solution to this issue is to include research ‘users’ in the assessment panel (see case studies below). However, identifying who these ‘users’ might be, and ensuring coverage of all the relevant applications and impacts of the research that might occur, is challenging. Similarly, engaging these users is also not straightforward, since they are not part of the academic system and hence have less motivation to engage in a peer review process which can be time consuming and typically is not accompanied by any financial reward.

Transparency is another possible criticism of peer review. Individual reviewers are often anonymous, and the decisionmaking process in review committees is intrinsically opaque to those outside. However, despite this possible limitation, acceptance of peer review is generally high, as described previously.

Finally, peer review tends to be slow. Decisions on a review process often take in excess of six months (and in some case over a year), which can prove a significant challenge in some circumstances, and delay the progress of the research process. This can be even worse when specific deadlines are required, for example in funding processes, which may result in a ‘glut’ of proposals as researchers rush to meet a deadline. Similarly, such deadlines limit the ability of researchers to respond to research needs as they arise. Thus peer review can limit the timeliness of research and its publication.

Peer review can be used at different levels of aggregation depending on the needs of the review. However, it is usually applied at the individual or grant level. At higher levels of aggregation, there is a challenge for reviewers in understanding the range of research being conducted – here peers are more difficult to identify, and reviewers are not subject experts in all the material reviewed, so the review is likely to be less academically rigorous.

Peer review can be combined with and informed by other approaches, including case studies and bibliometrics. Reviewers need to be presented with relevant information to review and compare. This may just be in the form of a set of academic publications, but could include other inputs drawing on wider methodologies.

**How is peer review used?**

Peer review can be used in a number of contexts. Most broadly, they can be separated into prospective assessment of research proposals before the work is conducted, generally in research funding applications, and retrospective evaluation of research which has been conducted. This can be in the assessment of articles for publication, or in wider assessments of institution or programme performance, for accountability or allocation purposes. In all of these cases, however, the approach used is broadly similar, corresponding to three key steps as outlined in the introduction, and explained in more detail below.
The review process

This step involves the review of materials by the peers identified. This usually takes place in one of two ways: either the materials are sent to individuals to review separately, or a group or panel of reviewers is gathered to review the material collectively (after having been given some time to review the materials in advance individually in most cases). Each of these approaches has advantages and disadvantages.

The advantage of a group or panel review is that it allows for discussion and clarification of any misunderstandings among the group, and gives a better opportunity for consensus to be reached. It is also time limited, in that decisions need to be reached by the end of the panel session. However, there is the risk that one or two individuals may dominate the process if it is not well moderated, and there is evidence that panels tend to be even more conservative than individual peer reviewers, as there is social pressure to conform in such a group dynamic. Furthermore, there are costs associated with bringing together reviewers in a particular location.

Using individual reviewers eliminates travel and venue costs, and the impact of the group dynamic on the review process. It also allows specific comments to be provided by individual reviewers focused on the materials provided. Comments from a group review may be more general, although this is not always the case. However, individual review approaches can increase the time to complete the review, as there is no time limiting session where decisions have to be made – there is evidence that only around half of reviewers complete their review in the allocated time in such approaches (Harding, 2002). Furthermore, there is the chance that consensus on the outcome will not be reached among reviewers. In this case, it may be necessary to bring in additional reviewers to resolve the conflict, and this can again be time consuming.

In many cases, the author is given the chance to revise and review their work in response to reviewer comments, if it is felt that this may improve the materials and increase their chance of a successful review outcome. This increases the time taken for the review process overall, but can be particularly beneficial for young researchers, who may be less experienced in preparing materials for review, or for more innovative research, as it allows further opportunity for explanation and justification where reviewers express concern.

Triage

Before reviewing, there is often a preliminary 'triage' stage, the purpose of which is to keep the peer review process manageable and reduce burden. Here, a preliminary assessment is made of the materials to be reviewed. The aim at this stage is not to make any assessment of the relative quality of the materials but to ensure they meet the basic standards for review. This includes a check on whether all the relevant materials and information for review have been provided. The process also includes ensuring that the materials meet any minimum criteria set for the review. For example, where funding applications are made in response to a particular funding call, this step may involve checking that the proposals received fall within the eligibility criteria of the funding scheme, in their topic, applicants, proposed budget and so on. This step may also include identifying the diversity of academic fields or topic areas included in the materials for review, and hence identifying and recruiting the relevant reviewers. In some cases, this reviewer recruiting process may have already taken place before the call was issued, for example where applications are received in response to a call for proposals in a very specific area.

Outcome

The final stage of a peer review process might be to make a yes or no decision, for example on whether a piece of research should be funded, or whether a paper should be accepted for publication. However, the outcome may instead be some form of score or ranking, which can then be used by the assessment body. This may be to provide a comparative ranking of different researchers, or (following aggregation if necessary) institutions, or the assessment body may use the scores to determine a cut-off point for funding of proposed research, depending on the funds it has available. Where scores come from a number of individual reviewers, the assessing body needs to devise an appropriate method for combining these scores. In some cases, though less frequently, the outcome may be a qualitative review by the review-
ers or panel of the overall performance or quality of a particular grant, researcher, group or institution. This may include feedback, for example on areas for improvement, or recommendations. This type of outcome is most typical in national assessments of the research base, such as in the Standard Evaluation Protocol in the Netherlands.

Although the above process illustrates the typical steps involved in peer review, there are a number of modifications which can be made to the peer review process in order to match it better to the needs of different assessment systems. One of these, as described previously, is the use of interdisciplinary or multi-stakeholder review panels, to bring in different viewpoints beyond the pure academic merit of a proposed research project. Participants could include decisionmakers, community members and academics from wider subject areas. Another possible adjustment is to adapt the research funding process to reduce conservatism. This can be achieved in a number of ways. One method is to use a portfolio approach, with funds allocated with a balanced risk approach, rather than by topic. Another alternative is to reserve a portion of a funder’s money for ad-hoc or discretionary funding of projects that fall outside the boundaries of predefined programmes, to support unique, worthwhile work that may not otherwise be funded in a timely manner.

**Example case studies**

The example case studies selected all demonstrate the use of peer review in large-scale (national) assessments of performance and/or quality across institutions, reflecting the focus of this study. This does not reflect the broader use of peer review in research evaluation, which spans a wide variety of possible applications, as described previously.

**Excellence in Research for Australia**

In Excellence in Research for Australia (ERA), peer review is used to allocate a score to a research group or departments within institutions against a five-point scale, which is internationally benchmarked. The review is by panel, and on the basis of a range of indicators, which is provided to the panel to demonstrate the performance of the research group across a number of areas. Because of the nature of the information provided, the panel members are not just disciplinary experts, but also need to have some experience in research evaluation, as the process is largely to compare different indicators, rather than to read and compare different research outputs.

According to the Australian Research Council the expert panel should be made up of ‘internationally recognised members with expertise in research evaluation and with broad disciplinary expertise’ (2010). Indicators include citation data on publication outputs for some subjects, other quantitative data on research quantity, such as numbers of publications and research income, and indicators of wider societal impact, such as number of patents published, or citations on clinical guidelines. Having the expertise to deal with these kinds of indicators and understanding their limitations is therefore important in order to make an effective assessment of performance.

In addition to this expert review, more traditional forms of peer review are also used in ERA to assess the quality of research in subjects where bibliometric indicators are not proven or are ineffective – largely the arts and humanities. These reviewers can come from within the expert panel conducting the wider review, but the panel is able to bring in further subject experts where required if there is not appropriate expertise on the expert panel to conduct the assessment. When peer review is used, quality assessments are not made on individual outputs, but on the overall sample submitted for review, which will cover 20 per cent of all outputs. Overall, the ERA system is designed to try and minimise burden while still conducting a rigorous assessment. Thus the peer review of publications is light touch, and replaced by bibliometrics where feasible, and the expert panel is provided with a clear and comparable set of indicators, largely quantitative, to facilitate quick and clear decisionmaking.

**Research Excellence Framework**

In the Research Excellence Framework (REF), peer review is used in a wider context than typically, to assess not just academic excellence but also the wider impact of research and vitality of the research environment. Unlike ERA, there is no specific requirement for the panel members to have experience in
research evaluation. However, the subject-specific sub-panels which carry out the assessment are supported by four overarching main panels, which provide guidance and leadership, as well as ensuring consistency between the subject-specific panels. The panels have the option to bring in research users for the assessment of the impact of research. Although not all users are to be involved from inception, at least one research user is appointed to each sub-panel early enough to be involved in the development of assessment criteria.

The materials provided to the panels for the assessment differ for the different elements of the assessment. For the assessment of research quality, they are provided with a set of publications, with some additional bibliometric information in some cases. For the impact assessment, they are provided with a series of case studies and an overarching impact statement. For the assessment of research vitality, they receive a completed template outlining strategy, staffing and so on. All these materials are predominantly qualitative, and the work of the panel in dealing with these submissions may be significant. Assessment takes place at the group or department level within institutions. The outcomes for each unit are combined into an overall rating for institutions by the assessment body. The outcomes of the assessment process ultimately determine allocation of core funding to universities in the UK.

**Standard Evaluation Protocol**

The purpose of the peer review assessment in the Standard Evaluation Protocol (SEP) is somewhat different from the two previous examples, in that the objectives are accountability and improvement. In this context, the review is intended to be retrospective and prospective, reflected by the criteria used in the review. The evaluation committee is required to be independent of the institute, well acquainted with the current research practice of the discipline(s) and cover the various other areas of the institute’s activities (e.g. PhD training, research in the context of application, provision and maintenance of research facilities for other academic and non-academic target groups). Research management competence is to be represented in the committee (VSNU, KNAW and NWO, 2009).

The assessment covers quality, productivity, relevance, and vitality and feasibility at research group or subject area level and institute level. The assessment process includes a site visit and interviews, as well as the analysis of the documentary evidence provided in the self-evaluation report. This requires a different set of skills from those in peer review of, for example, a set of publications, and it is not clear whether any specific preparation or training is given to the committee members in advance, although guidance is provided by the assessing board (the Association of Universities in the Netherlands, the Royal Netherlands Academy of Arts and Sciences or the Netherlands Organisation for Scientific Research) that has appointed the committee. Clearly the committee is also selected with the needs of the review in mind and reflects a larger population than just relevant academic experts. The outcome of the assessment includes ratings of the performance of the university or institute, but the key outcome of the process is an evaluation report, which includes recommendations for improvement.

**Summary**

Peer review is frequently used in research evaluation for a range of purposes and is a highly respected method for assessing the quality of research. Although it can be burdensome, and may be conservative, it is widely applicable in several contexts, and is credible to the academic community. The approach is adaptable and has been used to assess not just quality, but also wider elements of research such as societal impact. In these cases, the approach can be strengthened by bringing in others, who may be from outside academia, with specific relevant expertise into the reviewing panel. Peer review can be readily combined with many other tools including bibliometrics, case studies and interviews.
Introduction

Site visits are primarily used to give evaluators the opportunity to interact with members of the department or institution being evaluated. They generally consist of meetings over one or more days between a panel of expert evaluators and members of the department. This can give evaluators the opportunity to communicate with a wide range of stakeholders across the department, and to set the findings and recommendations of the evaluation in context.

When should site visits be used?

One of the key advantages of site visits is that they give evaluators the opportunity to speak to several people at different levels. This is one of the main motivations for using site visits, as they give access to people who would not normally be consulted in other approaches, and the opportunity to canvas many viewpoints. Furthermore, it is an interactive process, giving institutions the opportunity to understand and contribute to the evaluation process. This enables the evaluators to engage with a wide range of the relevant stakeholders in the evaluation process, and by being more open and interactive, it creates a better sense of a two-way process rather than an externally imposed evaluation.

One of the key disadvantages of site visits is that it is expensive in time and costs to send the evaluation panel to each individual institution under assessment, and they may not be highly scalable. It would be difficult to implement an approach reliant on site visits across a large number of institutions. Furthermore, site visits can be burdensome for participants, as a large number of different stakeholders may be involved, and institutions may spend a long time in preparation for even a short visit if they feel that the evaluation process is highly dependent on the visit. Similarly, the time constraints mean that the evaluation is inevitably a snapshot. Site visits are not something that can be repeated frequently without prohibitive cost implications, and in each visit time is limited and it is only feasible to speak to a limited group of people since visits, tend to take place over the course of only one or two days. Also, the process is not usually to be transparent to those outside the institution (though it may be more transparent to the institutions participating than many other approaches) nor highly objective since it depends on the interaction between a number of individuals.

Site visits commonly take place at the institution or department level, but with input at lower levels of aggregation in interviews and focus groups. They tend to be used for validation in response to an initial self-assessment process, such as in SEP or AERES, and thus fit with a number of other tools, depending on what is used in the initial self-evaluation step.

How are site visits used?

Site visits are not typically a data gathering step, but a data validation and development process. Generally, site visits are most appropriate to use in formative evaluations, where having the opportunity to understand context and follow an interactive process is important. There is no fixed process for site visits, but they generally consist of a series of meetings with stakeholders, individually and in groups. This is usually followed by a session in which members of the evaluation panel discuss their findings and conclusions. Typically the site visit follows some ini-
Site visits are a tool that is often used in research evaluation to give evaluators and evaluated participants the opportunity to interact, and to allow a greater range of perspectives and participants to be taken into account in the evaluation process. They are best used in formative evaluations, to give context and a wider range of inputs to improve the quality of the advice and recommendations of the evaluation panel. They are generally used to validate and develop the input from a prior report, often the result of a self-assessment process.

Summary

Site visits are a tool that is often used in research evaluation to give evaluators and evaluated participants the opportunity to interact, and to allow a greater range of perspectives and participants to be taken into account in the evaluation process. They are best used in formative evaluations, to give context and a wider range of inputs to improve the quality of the advice and recommendations of the evaluation panel. They are generally used to validate and develop the input from a prior report, often the result of a self-assessment process.
Introduction

Surveys provide a broad overview of the current status of a particular programme or body of research and are widely used in research evaluation to provide comparable data across a range of researchers and/or grants which are easy to analyse. Surveys can be employed in a range of ways, including online, by post or by telephone. They tend to provide broad rather than deep information, but the information is comparable between respondents and it is possible to gather and analyse data across a wide range of people and/or groups.

When should surveys be used?

Surveys are used to collect information across a group of people, covering either a whole population or a selected sample. Information gathered can be quantitative or qualitative, and is often a mix of both, but usually providing an overview across the range of people, rather than specific details, as surveys are not appropriate for gathering in-depth qualitative information.

Limitations and challenges

There are various key potential limitations for surveys. As stated above, probably one of the main limitations is the depth of information gathered. Survey data tend to consist of broad rather than deep information, although this depends to some extent on the survey design. It is more difficult to use survey information to provide a comprehensive understanding of all the contextual information around a particular researcher’s work or the outputs of a particular research grant.

Another key limitation is in the adaptability and appropriateness of the method. As the methodology relies on standardisation across respondents, questions have to be general enough to cover all respondents, which limits their applicability and relevance to each individual respondent. Therefore issues that are very important, but only to a limited number of individuals, may well be missed. Similarly, the format puts a strain on the validity of the data collected. Where closed questions are used (and these are often desirable to facilitate comparability and ease of processing of data), respondents are forced into a fixed set of answers, which have to serve as proxies for the ‘real’ answer. For example, people’s real experiences are often hard to grasp in simple polarised answers such as ‘agree/disagree’. These are likely to serve only as approximations of what we intend when we compile the survey, and what the respondents mean by their response (CSU, n.d.).

Inflexibility is also a challenge in the use of surveys. The initial survey design usually remains unchanged throughout the period of data collection. Even where methods used make changes to the instrument possible, it is often undesirable because of the challenges in comparability that would result. This demonstrates the importance of appropriate piloting of survey tools.

As well as the intrinsic limitations described above, there are also a number of challenges which need to be addressed when using a survey, and should be taken into account when determining if this is the appropriate approach to use in a particular set of circumstances. The first is the resource requirements of the survey. Surveys can be a cost-effective way to reach a large number of respondents. However, the type of survey will have a significant
impact on the costs, as outlined in more detail in the following section.

One thing which can add significantly to the cost of a survey is the response rate required. In some cases, in research evaluation this problem can be minimised, for example where survey response is a requirement of a funding allocation. However, where response is not strongly incentivised, low response rates can undermine the representativeness of the survey results, and where high response rates are required, this can add significantly to the time and costs of administering a survey. As described, response rate is one factor which can influence the representativeness of the sample used, and if the survey is not to be an audit of the full population of interest, care needs to be taken to ensure the sample selected will give the information required.

This leads on to the specialised skills required for survey approaches to be employed effectively. Clearly, where sampling approaches are used (and even, to some extent, where they are not), survey approaches require some statistical knowledge, understanding of sampling approaches, and other specialised skills in order to process and interpret the results. Some survey approaches also require further specialised skills in delivering the survey. These could be interview skills (for face-to-face or telephone interviews) or IT skills (for online surveys).

**Strengths**

Despite these challenges, surveys have a number of strengths, which make them an important research tool. Perhaps most importantly, surveys are able to gather comparable information across a population. The standardisation of the approach facilitates this comparability. Surveys where mostly closed-ended questions are used give broadly quantitative results, which further aids comparability. Usually, the survey data produced is highly reliable. As the stimulus is standardised, bias and objectivity resulting from the role of the researcher is minimised, particularly in web or postal surveys, which also minimise the interaction between surveyor and respondent. Careful wording and construction of the question set minimises unreliability of the subject, and where interview surveys are used, bias from the researcher can be minimised with appropriate training of interviewers (Research Methods Knowledge Base, 2006).

Survey data can be generalised to the entire population provided the sampling approach is appropriate. Sampling can also potentially be avoided, as surveys can be used with large numbers of people (though this can have cost implications). In the context of research evaluation, surveys are likely to be a good approach when you want to gather information covering all researchers or research grants in a particular portfolio, as they allow standardised information to be gathered across a portfolio. This is because surveys are relatively inexpensive for the number of respondents that can be covered (especially self-administered surveys).

Another potential advantage of surveys is that they can be completed in a relatively short period of time, covering many respondents and gathering a significant amount of data. Similarly, data analysis can take significantly less time (particularly if largely closed-ended questions are used) than other qualitative approaches. Speed of deployment and analysis is particularly notable in the case of electronic surveys, which can be delivered across a population almost instantaneously.

**Level of aggregation**

Surveys are best implemented at individual or grant level. The survey approach generally requires a single respondent to have all the necessary information to respond to the questions included. The exception to this is focus groups, which are a survey method that allows responses to be made by a group and questions to be answered through collective discussion. However, this tends to be most relevant when perceptions and opinions are being gathered, which is less relevant to a research evaluation context, where the material to be collected is more likely to be of a specific factual nature.

**Cost or burden**

The cost or burden of survey approaches, centrally and for respondents, is strongly dependent on the type of survey employed. For example, a web-based survey with a limited set of closed-ended responses can be a low burden on all sides, whereas an open-ended interview (face-to-face or telephone) survey is likely to lead to a much higher burden and costs centrally, and may be also a greater burden to respondents. However, the
How are surveys used?

Types of surveys

There are two key types of survey: oral ‘interview’ surveys and written ‘questionnaire’ surveys (Figure A23). Questionnaires tend to consist of short closed-ended questions while interviews are most likely to contain broad open-ended ones, but not strictly divided this way.

Surveys can be delivered in different ways. Within oral interview surveys, there are two main groupings: telephone and face-to-face surveys. However, these boundaries are blurred by modern technology with video conference or Skype interview surveys also potential options. There are several advantages to interview survey approaches. First, it is possible to ensure that the appropriate person is responding and that the survey has not been passed on to be completed by another person, rather than the intended respondent. It is also possible for interviewers to document the characteristics of non-respondents and their reasons for refusal. This has a secondary advantage that response rates are usually higher, as it is possible for interviewers to alleviate any concerns about completion. Similarly, it also reduces non-response to individual survey items. Interview surveys also can be preferable in cases where surveys address complex issues and more detailed explanation is needed.

The key disadvantage of these approaches is cost, and this is an even more significant problem in face-to-face interviews than telephone interviews. Part of this cost is in the recruiting and training interviewers. As well as being expensive, the process is
time consuming, and again this is even more the case for face-to-face interviews. Another possible disadvantage is the potential for a social desirability bias affecting the outcome of the survey. This is when people adjust the answer to 'please' the interviewer, or to avoid offending them, or otherwise modify their response because of the social element related to a person-to-person interview. This can be minimised by appropriate training for interviewers (Research Methods Knowledge Base, 2006).

Similarly, questionnaire surveys can take a range of different forms, from postal surveys, to surveys conducted in person in a central location, and more recently electronic surveys. Again, there are a variety of advantages and disadvantages for each of those approaches. One of the key advantages is cost. Administrative costs and costs per respondent tend to be significantly lower than for interview surveys. Similarly, deployment time can also be an advantage, particularly in the case of electronic surveys, which can be deployed very rapidly. In addition, the social desirability bias is minimised in questionnaire surveys.

The key disadvantages are response rate and time. Although surveys are quick to deploy, it is hard to control when respondents will complete the survey, which can result in long delays awaiting responses. Similarly, response rates tend to be lower, and often it is not possible to determine the demographics and characteristics of non-respondents and their reasons for not responding. Finally, even where responses are received, some questions may not be completed, although this can be controlled in some cases in electronic surveys.

**Key choices and issues in survey design**

Although the method of delivery is one of the key choices in survey methodology there are a variety of other questions that need to be answered in order to establish the nature of the survey to be conducted, and which may have an influence on the method of delivery selected.

The structure of the survey needs to be clarified. One of the most crucial questions is whether questions will be open ended (without pre-specified responses), closed ended, or a mixture of both. This will depend on whether it is possible to anticipate the most frequent or important types of response and hence develop closed-ended questions. It will also depend on what type of analysis you wish to conduct on the survey data. Clearly closed-ended questions result in data which are more amenable to quantitative analysis.

Also important to consider is the likely complexity of the questions. For example, will you have questions with multiple parts or sub-questions, in which case it may be valuable to be able to control the question sequence so that some questions are only seen when specific answers to earlier questions are given. The extent to which this can be controlled depends on the format. For example, with an oral interview, it will be possible for the interviewer to construct the survey relevantly, whereas a paper survey delivered by post will obviously be more inflexible (even if respondents are instructed to skip certain questions).

Another possible issue is screening. Do you need to screen respondents to ensure they are able to answer the questions in the survey? For example, if you are asking for information on the way in which graduate students are engaged in research, you may want to screen to remove researchers who do not employ any graduate students.

Content issues are also important in survey design. For example, if the respondent needs to consult records or refer to other materials (e.g. a list of publications) it is likely that the survey might be best presented in a written format by post or electronically, as respondents may feel uncomfortable looking up details during a phone call or while waiting for a response during a face-to-face survey.

Feasibility considerations are also crucial. Clearly financial considerations influence the type of survey conducted, and whether the survey is an audit (covers the whole population) or a survey of some form of sample of the population. This also depends on the size of the full population. In the context of research evaluation, audits often are feasible and indeed necessary for the needs of the evaluation. However, if they are carried out across a large group, it may make interview survey approaches unfeasibly costly. Similarly, the timescale over which responses are required also influences the approach selected. For example, electronic surveys can be distributed much more quickly than a survey distributed by
post. In interview surveys, personnel issues are also important. Interviewers need thorough training if the survey is to be delivered correctly and without biasing the response.

**Sampling**
If the survey will not be an audit, and will select a subset or sample of the full population of interest, sampling approaches need to be taken into consideration. Here there are two key considerations: what will the size of the sample be, and how will the sample be selected? The answer to the first of these questions largely depends on the time and budget available, but also needs to take into account the confidence you need in the results produced, and the size of the target population you are investigating. Sample selection requires further considerations. The first is whether an audit (a survey covering the whole of the target population) is possible and desirable. In research evaluation this is often the case, and sample selection considerations are not significant. However, if the target population is large, and sampling is necessary, there are several approaches which can be taken.

The first is to conduct a random sample, in which any member of your target population has an equal probability of being selected. This requires some careful thought to ensure that this is truly the case. For example, if you are conducting a web survey, do all members of your population have internet access? An alternative approach is to select what is termed a convenience sample, which is just to select a sample that is easily accessible to you. This reduces generalisability to wider populations, but may still be appropriate in some circumstances, for example where generalisability is not as important as ensuring you have certain individuals included in your survey. Other sampling approaches exist, many of which are adaptations or developments on the random sampling approach designed to allow for some of the challenges in ensuring the sample is truly random, and the difficulty in accessing certain groups that may fall within your population (Research Methods Knowledge Base, 2006).

In research evaluation, it is not usually difficult to access participants, as most researchers and research managers are well known to surveying bodies, so if an audit cannot be conducted, a random sample is usually feasible. However, there are situations where the issues around ensuring a sample is random may occur, for example, when attempting to identify research users and other stakeholders outside the academic community. Similarly, when studying research retrospectively, contacting your population may be challenging if researchers have moved on and potentially left academia or moved institution since the research you are investigating was conducted.

**Step-by-step guide**
Taking into account the above considerations, we can define a series of steps necessary to conduct a survey. These steps assume that initial decision making has taken place and that it has been concluded that a survey is the most appropriate approach to use in this circumstance, on the basis of some of the factors outlined in the previous section. Once that decision has been made, the survey process follows the following steps, based on the Health Communication Unit at the University of Toronto (1999):

1. **Decide what the purpose of the survey is:** The first step is to establish the purpose of survey clearly. Key questions here include: What is the population of interest? What issues and questions need to be explored?
2. **Find out what resources are available:** The next key step is to understand what resources are available and can be accessed for the survey. This can include personnel as well as financial resources.
3. **Decide on methods:** After determining the purpose of the survey and the available resources, the next step is to decide which methods are most appropriate, taking into account the range of considerations outlined above. The sample design forms part of this step.
4. **Write survey:** The next step is to design the survey. Considerations include not only the questions to ask, but also the order in which to present them, whether questions will be conditional on previous responses, the types of response formats that will be used, and for questionnaires the layout.
5. **Pilot test and revise the survey as necessary:** This step is important in ensuring the survey is appropriate and to deal with any issues before the
survey is deployed, as it is difficult to revise the survey once it is deployed because of the difficulty of maintaining consistency.

6 Select the sample: The overall sampling strategy needs to be considered when deciding on the survey methods in step 3. This is the point at which the sample has to be identified and selected for deployment.

7 Train interviewers (if necessary): Depending on the approach selected, it may be necessary to recruit and train interviewers. This can be time consuming and is important in ensuring a successful interview survey, so sufficient time should be allowed for this step.

8 Collect data: This stage is the deployment of the survey.

9 Process data: The extent to which data processing is time consuming depends on the survey approach used. This may consist of data entry, and can include coding and interpretation where open-ended questions are used.

10 Analyse the results: Finally, the data can be analysed, using appropriate statistical approaches, and final results generated and interpreted.

Example case studies

The RAISS survey

One of the survey approaches which has been developed at RAND Europe is the RAND ARC Impact Scoring System (RAISS) tool (Wooding et al., 2009), which has been used to evaluate a broad range of R&D activities. This survey was originally developed for the Arthritis Research Campaign (ARC, a large biomedical research funder in the UK), where it replaced their end-of-grant reporting system, and was subsequently refined and adapted for the National Institute of Health Research (England), which rolled out the system to over 3,000 researchers it funds (Starkey et al., forthcoming). It has now been refined for application in other sectors outside health, building on previous surveys conducted in other research fields.

The RAISS tool is a web survey and one of the key considerations in its design was to minimise burden on participants. By being quick to complete, participants are not over burdened, leading to high response rates, which allows the tool to be employed in a wider variety of ways. For example, several funders have used the tool to look at the outputs and impacts of research across the lifetime of funding programmes, rather than in a limited ‘one-off’ way after the research projects are finished. Response rates in any case have not generally been an issue as the survey has largely been used by funders to collect data from their grant holder, in which case the surveying body has significant leverage on the survey participants.

The RAIISS survey consists of 187 Yes/No questions that explore what impacts a piece of research has contributed to. By using high-level questions to determine which more detailed questions are asked and by asking exclusively Yes/No questions, the average completion time of the questionnaire is kept to just over half an hour, with 90 per cent of questionnaires completed in less than an hour. However, the questionnaire still collects information across a huge range of impacts, including career development, collaboration within and outside academia, research capacity building, research tool production, dissemination in academic and non-academic contexts, impacts on health policy through a variety of routes, impacts on education and training of others, and a wide variety of intervention, product development and public health advice outcomes.

The selection of such simple closed-ended questions was intended not just to simplify the questionnaire for respondents, though this was important, but also to facilitate quick and easy analysis of the data produced, and to enable data to be represented graphically. The outputs of the tool can be represented as an impact array that gives an instant overview of the research portfolio. Figure A24 shows an example of an impact array, in which:

- the columns represent different types of outputs and impacts from grants or projects
- the rows represent different projects or grants in a portfolio
- the colouring represents the intensity or geographic reach of an impact.

These types of displays can also be used to identify different process variables (e.g. types of collaboration structures) and to then look into correlations with performance.
Strategic planning (10 per cent): to assess whether the agency sets valid annual milestones and long-term goals for the programme. (Sample questions: Does the programme address a specific and existing problem, interest or need? Is the programme designed so that it is not redundant or duplicative of any other federal, state, local or private effort?)

Programme management (20 per cent): to rate agency management of the programme, including financial oversight and programme improvement efforts. (Sample questions: Does the programme use strong financial management practices? Does the programme collaborate and coordinate effectively with related programmes?)

Programme results (50 per cent): to rate programme performance on goals reviewed in the strategic planning section and through other evaluations. (Sample questions: Has the programme demonstrated adequate progress in achieving its long-term performance goals? Does the programme demonstrate improved efficiencies or cost-effectiveness in achieving programme goals each year?)
In contrast to RAISS, the questions are open ended, with respondents providing, for each question, a short answer plus a more detailed explanation with supporting evidence. These are reviewed and a rating assigned centrally, so although there are only 25 questions included, this is not necessarily a low burden approach at either end. Despite this, the response rate is not likely to be an issue given the influence of the surveying body on the programmes surveyed.

The survey is deployed at a programme level, and although not all federal programmes were covered in the first instance, a sampling approach was not necessary since this was an audit, over the long term, and since the primary intention was to assess programmes individually and against their own goals, rather than comparatively. Nonetheless, some comparison is inevitable and indeed the number of programmes achieving different rating levels was outlined on the associated website (http://www.expectmore.gov), intended to provide transparent information on the performance of federal programmes.

Impact of the Health Technology Assessment Programme

A study published in 2007 used a postal survey as part of a wider study into the impact of the Health Technology Assessment (HTA) Programme funded by the NIHR in the UK (Hanney et al. 2007). The study also used case studies and document review to investigate the impact of the programme, as well as drawing on the ‘payback framework’ as a logic model to understand the research translation process.

The survey sample covered researchers who had been funded through the programme over a certain period and who had submitted a final report. There were some exclusions where projects had been discontinued, or researchers were not available for survey. This is an example where the sample can be virtually a full population as the population size is relatively small, which is not atypical in many research evaluation scenarios. The survey built on the logic model used for the study, the payback framework, which provided a structured way to build the question set and analyse the results. Some initial efforts were taken to phone respondents and check their addresses, or find second author respondents where first authors were not available, which helped boost response rates. A postal rather than email survey was used as the central body was not willing to share email addresses with the research team. Various techniques, including personalisation, provision of a stamped return envelope, and ensuring the relevant university logo was visible on opening of envelopes were used to promote response.

The survey was sent a further two times to those who did not respond. The final response rate was around 65 per cent. The researchers found that a significantly higher number of outputs from research, including publications, conference proceedings, presentations and so on, were identified through the survey than had previously been identified through end-of-grant reports. It may have been possible to collect some of these data, such as peer reviewed publications, through internet searches or bibliometrics. However, much data, such as information on presentations or follow-on research, could only be collected comprehensively from the relevant researchers.

Summary

Surveys can be used to provide an overview of information across a wide range of respondents. The key limitation of this research approach is the lack of detail and specificity in the data produced, and key strengths are the ability to cover a large number of people, and cost- and time-effectiveness. Challenges in running a survey include ensuring the necessary response rate and the need for specialised skills centrally, such as trained interviewers, or people with IT skills to deploy web surveys. In research evaluation, surveys tend to be used to provide an overview of a research programme or portfolio, and tend to be used to collect specific facts or data which are likely to be available for all grants or researchers. As suggested, surveys are best applied at the individual researcher or grant level, and in most cases web surveys are used, for simplicity and because of their low cost. Surveys can be used with other techniques, such as case studies, if more detailed qualitative information is needed to support the overview provided by the survey.
What is visualisation?

Data visualisation takes advantage of humans’ capacity to process information and extract meaning through visual representations. It is a tool for data summarisation: it involves taking large amounts of data and presenting them in a visual format for human comprehension and interpretation. Visualisation tools are becoming increasingly important, given the ever expanding amount of data available for analysis; one estimate from the University of Berkeley is that there is approximately one exabyte of data generated annually (Keim, 2002).

Data visualisation is often used in combination with data mining; while data mining refers to processes for identifying patterns and insights from data, visualisation can be a tool for presenting these data (Kohavi, 2000). Also, visualisation can be used instead of automated data mining techniques, for example, by presenting data in a way that allows them to be interpreted visually. This is of particular use when data are highly heterogeneous and noisy; it also allows data to be explored intuitively without the use of mathematical or statistical algorithms.

What type of data are appropriate to visualisation?

There is a wide breadth of techniques available for data visualisation, which can be applied to existing datasets, involving multiple variables and dimensions. However, the challenge is in selecting and using techniques to ensure an accurate representation of the data, which also highlights the features of the data of interest for the analysis. This requires a clear understanding of the nature of the data and the objective of the analysis.

There are two main categories of visualisation techniques:

- **Data exploration techniques**: These include standard 2D and 3D techniques (eg bar charts, line graphs), geometrically transformed displays (eg scatterplot matrices, parallel coordination technique), iconic displays (eg mapping data to ‘icons’, such as star icons, stick figures, colour icons, tile bars), dense pixel displays (mapping values to coloured pixels) and stacked displays (presentation of partitioned data in a hierarchical fashion).

- **Data interaction and distortion techniques**: These allow researchers to change the visualisation dynamically and also focus on particular details, while still preserving an overview of the data. Examples include interactive filtering, interactive zooming, and interactive linking and brushing (eg multiple scatterplots, bar charts, parallel coordinates, pixel displays, maps).

Figure A25 illustrates some data visualisation techniques, and the type of data that can be visualised through them.

How can data visualisation be applied to research evaluation? Key examples

Data visualisation can be used in combination with automated data mining in research evaluation to present data from algorithms in a meaningful way, or instead of automated data mining to allow for visual analysis of the data.
Figure A25
Classification of information visualisation techniques

Source: Keim (2002)

Historically, data visualisation for science policy has been used in the field of bibliometrics, to build science or bibliometric maps based on publication data, for example, clustering documents based on co-citation or co-occurrence of author names. Bibliometric maps are scaled representations of clusters of papers, which illustrate a scientific field provided by the papers and the links between them (Buter and Noyons, 2003). Börner, Chen and Boyack (2003) suggest key steps in developing bibliometric or science maps, including data extraction, unit of analysis, measures, similarity method (e.g. co-occurrence, matrix type), ordination (e.g. reduction technique, clustering, scaling) and display. There has been a recent surge of interest in the potential applications and uses of bibliometric maps: new graphic interfaces, e.g. HTML, Java applets and other web interface techniques, provided an impetus for increased interest and use in science mapping. Interactive bibliometric maps (e.g. iBEX), which can retrieve and search for different combinations of statistics and other data, have also been developed in the past decade.

Another recent use of data visualisation for research evaluation is the RAIISS tool, developed by RAND and ARC. The RAIISS tool for research evaluation uses data visualisation for visual analysis of data. RAIISS is an information gathering tool that can provide a strategic overview of a funder’s research portfolio using a web-based survey instrument. Data gathered through the survey instrument are visualised in the form of an ‘impact array’ (Figure A24), using pixel colour shading to illustrate
progression from early stage to late stage impacts, or from local to international impacts.

A final example of the application of data visualisation to research evaluation is STAR METRICS. While still in the development phase, STAR METRICS is planned to use a form of data visualisation, to present the results of the automated data mining process. However, this is in its early stages and exactly how data will be visualised with STAR METRICS remains unclear.
Appendix E: Information on other reports and frameworks reviewed and consulted

In addition to the frameworks investigated in this study, we widened our final analysis to cover a number of research evaluation frameworks which have been investigated by RAND Europe in previous studies. We have not revisited these frameworks in this study, but used our previous analyses to score them against the characteristics identified to provide a larger dataset for our analysis. The methodology used in the analysis, and the need for additional frameworks to be included, is explained in Appendix B. These are the additional frameworks included in the analysis, with a note of where more detail about their use can be found:

- **DIUS**: the Economic Impacts in Research and Innovation Framework of the Department of Innovation, Universities and Skills (UK); more information available in Brutscher, Wooding and Grant (2008)
- **ERiC**: Evaluating Research in Context (Netherlands); more information available in Grant et al (2009)
- **EU**: evaluation system for Framework Programme 7 of the European Union; more information available in Brutscher, Wooding and Grant (2008)
- **LUMC**: Leiden University Medical Centre framework (Netherlands); more information available in Brutscher, Wooding and Grant (2008)
- **MORIA**: Measure of Research Impact and Achievement (Australia); more information available in Brutscher, Wooding and Grant (2008)
- **PART**: Programme Assessment and Rating Tool (US); more information available in Brutscher, Wooding and Grant (2008) and Grant et al. (2009)
- **RAISS**: RAND Assessment Impact Scoring System (UK); more information available in Grant et al. (2009)
- **RQF**: Research Duality Framework (Australia); more information available in Grant et al. (2009)
- **Vinnova**: Swedish governmental agency for innovation systems (Sweden); more information available in Brutscher, Wooding and Grant (2008).

As of 6 August 2012 the two reports listed here were available for free download from the RAND website at the following addresses:

Interest in and demand for the evaluation of research is increasing internationally. Several factors account for this: agendas for accountability and good governance and management, and fiscal austerity in many countries. There is a need to show that policymaking is evidence based and, in the current economic climate, to demonstrate accountability for the investment of public funds in research. This is complemented by a shift in the type of evaluation needed: from the traditional, summative, assessment to more formative evaluations and those covering wider outputs from research.

Given this growing need for effective and appropriate evaluation of research, it is increasingly important to understand how research can and should be evaluated in different contexts and to meet different needs.

This report provides a guide to the key considerations in developing an approach to research evaluation. It outlines the trade-offs that have to be taken into account and the contextual factors that need to be understood, drawing on experience of international approaches to research evaluation. In addition, a detailed overview of six research evaluation frameworks is provided, along with a brief overview of a further eight frameworks, and discussion of the main tools used in research evaluation. The report is likely to be of interest to policymakers, research funders, institutional leaders and research managers.

This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.