



EUROPE

THE ARTS
CHILD POLICY
CIVIL JUSTICE
EDUCATION
ENERGY AND ENVIRONMENT
HEALTH AND HEALTH CARE
INTERNATIONAL AFFAIRS
NATIONAL SECURITY
POPULATION AND AGING
PUBLIC SAFETY
SCIENCE AND TECHNOLOGY
SUBSTANCE ABUSE
TERRORISM AND
HOMELAND SECURITY
TRANSPORTATION AND
INFRASTRUCTURE
WORKFORCE AND WORKPLACE

This PDF document was made available from www.rand.org as a public service of the RAND Corporation.

[Jump down to document](#) ▼

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world.

Support RAND

[Browse Books & Publications](#)

[Make a charitable contribution](#)

For More Information

Visit RAND at www.rand.org

Explore [RAND Europe](#)

View [document details](#)

Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND PDFs to a non-RAND Web site is prohibited. RAND PDFs are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see [RAND Permissions](#).

This product is part of the RAND Corporation documented briefing series. RAND documented briefings are based on research briefed to a client, sponsor, or targeted audience and provide additional information on a specific topic. Although documented briefings have been peer reviewed, they are not expected to be comprehensive and may present preliminary findings.

DOCUMENTED
BRIEFING



Capturing Research Impacts

A review of international practice

Jonathan Grant, Philipp-Bastian Brutscher,
Susan Ella Kirk, Linda Butler, Steven Wooding

Prepared for the Higher Education Funding Council for England

The research described in this report was prepared for the Higher Education Funding Council for England.

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2010 RAND Corporation

Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Copies may not be duplicated for commercial purposes. Unauthorized posting of RAND documents to a non-RAND Web site is prohibited. RAND documents are protected under copyright law. For information on reprint and linking permissions, please visit the RAND permissions page (<http://www.rand.org/publications/permissions.html>).

Published 2010 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: <http://www.rand.org>
RAND Europe URL: <http://www.rand.org/randeurope>
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

In February 2009, the Higher Education Funding Council for England (HEFCE) commissioned RAND Europe to review approaches to evaluating the impact of research as part of their wider work programme to develop new arrangements for the assessment and funding of research – referred to as the Research Excellence Framework (REF). The objectives were:

- to review international practice in assessing research impact
- to identify relevant challenges, lessons and observations from international practice that help HEFCE develop a framework for assessing research impact.

The report presents the findings of our review, based on four case study examples of impact evaluation approaches: the Australian RQF, the UK ARC Scoring method, the US PART framework and the Dutch ERiC framework.

RAND Europe is an independent not-for-profit policy research institution that helps improve policy and decision making through research and analysis. For more information about RAND Europe or this document, please contact:

Jonathan Grant
President
RAND Europe
Westbrook Centre, Milton Road
Cambridge CB4 1YG
United Kingdom
Email: jgrant@rand.org
Tel: 01223 353329
www.randeurope.org

Acknowledgments

The authors are grateful for the interest and assistance of all those who contributed to this report. In particular, they would like to acknowledge the valuable input provided by Quality Assurance reviewers Sharif Ismail and Dr Emmanuel Hassan.

List of Abbreviations

ARC	UK Arthritis Research Campaign
DAG	Development Advisory Group
DEST	Department for Education, Science and Training
EAG	Expert Advisory Group
ERA	Excellence in Research for Australia
ERiC	Evaluating Research in Context framework
HEI	Higher Education Institutions
IRUA	Innovation Research Universities of Australia
KNAW	Royal Netherlands Academy of Arts and Sciences
MRC	Medical Research Council
NHMRC	National Health and Medical Research Council Australia
ODGT	Outputs Data Gathering Tool
OMB	US Office of Management and Budget
PART	US Program Assessment Rating Tool
RAE	Research Assessment Exercise
RAISS	RAND/ARC Impact Scoring System
REF	Research Excellence Framework
REPP	Research Embedment Performance Profile
RQF	Research Quality and Accessibility Framework
SEP	Dutch Standard Evaluation Protocol
SIAMPI	Project studying the Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions between science and society
WGRI	Working Group on Research Impact

Contents

PREFACE	I
ACKNOWLEDGMENTS	III
LIST OF ABBREVIATIONS	V
CONTENTS	VII
EXECUTIVE SUMMARY.....	IX
1. INTRODUCTION	3
2. CASE STUDY – RESEARCH QUALITY FRAMEWORK.....	7
3. CASE STUDY – RAND/ARC IMPACT SCORING SYSTEM.....	21
4. CASE STUDY – PROGRAM ASSESSMENT RATING TOOL.....	33
5. CASE STUDY – EVALUATING RESEARCH IN CONTEXT.....	47
6. ISSUES AND IDEAS – APPLYING LESSONS FOR THE REF	60
APPENDIX A: LONG LIST OF FRAMEWORKS	70
APPENDIX B: REFERENCES	77

Executive Summary

The purpose of this report is to inform the Higher Education Funding Council for England's (HEFCE's) formulation of an approach to assess research impact as part of the proposed Research Excellence Framework (REF). HEFCE has identified several criteria that would be significant in developing an impact assessment framework. The framework should:

- be credible and acceptable to the academic as well as user communities
- encompass the full range of economic, social, public policy, welfare, cultural and quality-of-life benefits
- within a single broad approach, be adaptable to apply to all disciplines
- be practicable and not generate an excessive workload for the sector
- avoid undesirable perceptions and incentives
- complement other funding streams including the research councils' approach to increasing the impact of research.

Methods

To inform their thinking, HEFCE commissioned RAND Europe to undertake an international review of how other research agencies measure impact. The objectives of the review were:

- to review international practice in assessing research impact
- to identify relevant challenges, lessons and observations from international practice that will help HEFCE develop a framework for assessing research impact.¹

Following a quick scan of international examples of impact frameworks, we selected four frameworks for further analysis: the Australian Research Quality and Accessibility Framework (RQF), the UK RAND/ARC Impact Scoring System (RAISS). Impact Scoring System, the US Program Assessment Rating Tool (PART) and the Dutch Evaluating Research in Context (ERiC).

A short summary of each framework is provided in Table 1. Details of the frameworks are provided in chapters 2–5.

¹ Please note the project was not meant to review the academic literature in the field.

Table 1 Frameworks reviewed

Framework	Country	Summary Description
Research Quality and Accessibility Framework (RQF)	Australia	<ul style="list-style-type: none"> ▪ The RQF is a worked up example of measuring impact in a higher education context. ▪ It takes the form of a case study approach: research groupings submit examples of research with high impact (and provide relevant evidence). ▪ The RQF was tested and is applicable to capture impact, but has never been implemented due to a change of government.
RAND/ARC Impact Scoring System (RAISS)	UK	<ul style="list-style-type: none"> ▪ The RAISS framework is a first step towards an indicator-based approach. ▪ It takes the form of a questionnaire (to be filled in by researchers) to capture over 150 potential research impacts. ▪ The framework has been used to capture the impact of research grants of the Arthritis Research Campaign (ARC).
Program Assessment Rating Tool (PART)	US	<ul style="list-style-type: none"> ▪ The PART framework is a self-evaluation approach used to assess programme performance across federal government. ▪ It takes the form of a questionnaire. It asks programmes to assess themselves against their own strategic (impact) goals. ▪ PART is used to assess impact and efficiency of over 1000 federal programmes.
Evaluating Research in Context (ERiC)	Netherlands	<ul style="list-style-type: none"> ▪ ERiC is a new framework to assess research impact (or 'societal quality') in the Dutch higher education system. ▪ It combines several evaluation approaches: ie self-evaluation; an indicator-based approach; and stakeholder analysis. ▪ ERiC is supposed to be integrated into the current higher education evaluation system this year.

The four frameworks were selected in order to ensure that the main evaluation approaches that are used in practice were captured: case study approach (RQF), indicator approach (RAISS), self-evaluation approach (PART) and a mixture of the three (ERiC).

The case studies were then discussed in a workshop with the HEFCE project team. The aim of the workshop was to identify and capture the main challenges, lessons and ideas that arise from the frameworks.

Also during the workshop, the HEFCE team and RAND Europe team assessed the four frameworks using a traffic-lighting methodology against HEFCE's criteria of an adequate framework – which are that it should be credible and acceptable to the academic as well as user communities; encompass the full range of economic, social, public policy, welfare, cultural and quality-of-life benefits, etc.

Findings

The key observations from the traffic-lighting exercise were:

- the work of the Australian RQF Working Group on Impact Assessment provides a promising basis for developing an impact approach for the REF
- impact indicators (as in the RAISS framework) are not sufficiently developed and tested to be used to make funding decisions.

The main observations in terms of challenges, lessons and ideas that arose from the workshop are outlined below.

Strategic intent

- To be effective it will be essential that the approach can provide an unambiguous rating and comparison of submissions.
- HEFCE could consider allowing some flexibility in the weights given to the three areas of assessment – quality, impact and environment.

- Because the REF will drive the behaviour of HEIs, it will be important to be very explicit about the criteria for assessing impact.

Defining impact

- Because of the diverse nature of impacts, HEFCE should consult widely and then provide a definition or statement of what it defines as impact.
- Because of the imperfections of both quantitative and qualitative measures, HEFCE should use a combination of the two, such as using case studies or narratives supported by possible proxy indicators of impact as proposed for the RQF.

Unit of analysis

- HEFCE should consider providing flexibility in how submissions are organised. This could be through the RQF approach which allows researchers and institutions discretion in how they organise into groupings – while still requiring groups to have some minimum size.

Attribution

- There is a need to acknowledge that the issue of attribution is complex, but can be mitigated by a case study approach, and by focusing on those impacts where attribution is more readily demonstrated. It needs to be made explicit that case studies are illustrative, rather than intended to measure the total impact of an HEI.
- A verification mechanism is likely to be important in building confidence in the approach.

Burden

- HEFCE will need to undertake further work to understand the likely burden of assessing impact, and compare it against the burden of previous RAEs.

End users

- Because of the key role of end users, HEFCE needs to do further work to understand the motivations, likelihood and commitment of end users in participating in the assessment of research impact.

1. Introduction

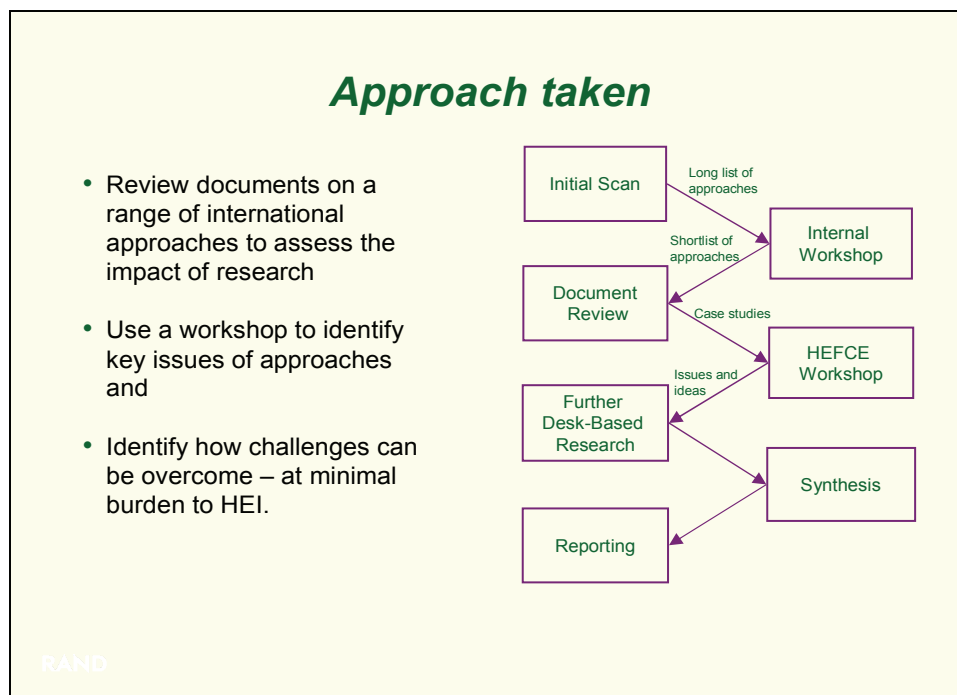
The Higher Education Funding Council for England (HEFCE) is working to develop new arrangements for the assessment and funding of research. HEFCE is working on behalf of the Scottish Funding Council, Higher Education Funding Council for Wales and the Department for Employment and Learning (Northern Ireland), and with a wide range of stakeholders to develop the new Research Excellence Framework (REF), which will be fully implemented by 2014.

The REF will consist of a single unified framework for the funding and assessment of research across all subjects. It will make greater use of quantitative indicators in the assessment of research quality than the preceding Research Assessment Exercise (RAE), while taking account of key differences between different disciplines.

Research assessment will combine quantitative indicators – including bibliometric indicators wherever these are appropriate – and expert/peer review. Which of these elements are employed, and the balance between them, will vary as appropriate to each subject. The economic and social impact of research, as well as impact upon public policy, will also contribute to the overall assessment of quality.

In this report we focus on the latter aspect – which is evaluating the impact of research. We have structured the report as an annotated briefing. The key messages of the report can be found in the slides at the top of each page, with text providing further details underneath. This layout highlights the major thrusts of the project through the slides, but also allows the reader the option to delve deeper into the detail of the findings as circumstances permit.

The report is divided into three sections. The introduction sets out the approach. The second section reviews the four case study frameworks in detail and the final section identifies some common issues and ideas that arise from analysis across the four case study frameworks.



The slide above outlines the approach employed for the study. We began with a quick scan of existing approaches for assessing research impact. This drew from our experience and knowledge of the field, existing work (Brutscher, Wooding and Grant, 2008) and web searches.

We collated summary information about a long list of frameworks for assessing research impact (listed on the next slide). We then selected four frameworks (as described below) for further review: the Australian Research Quality and Accessibility Framework, (RQF) the UK RAND/ARC Impact Scoring System (RAISS), the US Program Assessment Rating Tool (PART) and the Dutch Evaluating Research in Context (ERiC).

For each of the four frameworks we collated further information by reviewing available documentation, the web and speaking to or emailing key informants. We structured the information around a standard template in order to ensure comparability across the case studies – this included: what the framework was; why it was being used; how it was being used; who were the key stakeholders; and what were the consequences arising from the assessment?

This information is provided as case studies in Chapters 2–5 of this report and was presented to a workshop of the RAND Europe project team and HEFCE’s REF project team. The aim of the workshop was to identify and capture the challenges, lessons and ideas that arise from the research assessment frameworks.

Also during the workshop, the HEFCE project team and the RAND Europe team assessed the four frameworks against HEFCE’s criteria of an adequate impact framework – which are: to be credible and acceptable to the academic as well as user communities; to encompass the full range of economic, social, public policy, welfare, cultural and quality-of-life benefits; to be adaptable to all disciplines; to be practicable and not generate an excessive workload; to avoid undesirable perceptions and incentives; and to complement other funding streams including the research councils’ approach to increasing the impact of research.

The findings from the workshop are summarised in Chapter 5.

We selected four case study frameworks from a long list of 14 for further analysis

Reduced long list to a short list:

- Australia – Research Quality Framework (RQF)
- Australia – Excellence in Research for Australia (ERA)
- Australia – Measurement of Research Impact and Achievement (MORIA)
- UK – Department of Innovation, Universities and Skills (DIUS) Economic Impacts
- UK – Library House Model
- UK – Arthritis Research Council (ARC) scoring
- US – Congressionally Directed Medical Research Program (CDMRP)
- US – Program Assessment Rating Tool (PART)
- Netherlands – Standard Evaluation Protocol (SEP)
- Netherlands – Leiden University Medical Centre (LUMC)
- Netherlands – Evaluating Research in Context (ERiC)
- Sweden – Swedish Governmental Agency for Innovation Systems (VINNOVA)
- Canada – Payback
- Japan – National Institute for Academic Degrees and University Education (NIAD-UE).

In a quick scan we identified a long list of 14 frameworks (listed in the slide above and in more detail in Appendix A). The list is not comprehensive – but reflects what we could identify in the time available.

From the long list of frameworks we selected four frameworks for further analysis:

- The Australian Research Quality and Accessibility Framework (**RQF**) is a worked up, tested and applicable approach to capture research impacts, although it was not implemented due to a change of government.
- The UK RAND/ARC Impact Scoring System (**RAISS**) approach uses a questionnaire to capture over 150 potential research impacts and is a first step towards an indicator-based approach. The framework has been used to capture the impact of research grants.
- The US Program Assessment Rating Tool (**PART**) is a self-evaluation approach used to assess programme performance across federal government. PART is used to assess impact and efficiency of over 1000 federal programmes.
- The Dutch Evaluating Research in Context (**ERiC**) is a new framework to assess research impact (or ‘societal quality’) in the Dutch higher education system. It is meant to be integrated into the current higher education evaluation system this year.

The four frameworks were selected to ensure that the main evaluation approaches that are used in practice were captured: case study approach (RQF), indicator approach (RAISS), self-evaluation approach (PART) and a mixture of the three (ERiC). When more than one framework used a certain approach we chose the framework best suited in terms of likely transferability and scalability to the UK context.

<h2>Summary</h2>				
	What is it?	How is it being used?	Strengths	Weaknesses
RQF	<ul style="list-style-type: none"> ▪The RQF is the most worked up example of measuring impact in a higher education context. ▪It takes the form of a case study approach: research groupings submit examples of research with high impact. 	<ul style="list-style-type: none"> ▪The RQF was intended to inform funding decisions. ▪It was expected to (initially) determine c 10 percent of funding. 	<ul style="list-style-type: none"> ▪Credible and acceptable. ▪Encompasses full range of impacts. ▪Adaptable to all disciplines. 	
RAISS	<ul style="list-style-type: none"> ▪The ARC framework is a first step towards an indicator based approach. ▪It takes the form of a questionnaire to capture over 150 potential research impacts. 	<ul style="list-style-type: none"> ▪ARC scoring is not intended to inform individual funding decisions. ▪It is, however, meant to inform more general strategic decisions. 	<ul style="list-style-type: none"> ▪None in the context of the traffic light exercise. 	<ul style="list-style-type: none"> ▪Does not encompass full range of imp. ▪Difficult to adapt to all disciplines.
PART	<ul style="list-style-type: none"> ▪The PART framework is a self-evaluation approach to assess programme performance across federal government. ▪It is used to assess impact and efficiency of over 1000 federal programmes. 	<ul style="list-style-type: none"> ▪PART is supposed to inform funding decisions. ▪It is however only one of a number of factors influencing the level of funding. 	<ul style="list-style-type: none"> ▪Adaptable to all disciplines. 	
ERIC	<ul style="list-style-type: none"> ▪ERIC is a new framework to assess research impact in the Dutch HE system. ▪It combines self-evaluation, indicator based approach and stakeholder analysis. 	<ul style="list-style-type: none"> ▪The assessment is unlikely to have a direct impact on funding decisions. ▪Rather findings are intended to guide improvement. 	<ul style="list-style-type: none"> ▪Encompasses full range of impacts. 	<ul style="list-style-type: none"> ▪Likely to generate excessive workload.

The slide above provides a brief summary of the four frameworks that were studied in more depth. It illustrates the range of different assessment approaches employed, how the different frameworks make use of the assessment results (for example, to inform funding decisions) and the main strengths and weaknesses of the four frameworks.

The strengths and weaknesses are based on the assessment by the HEFCE project team and RAND Europe team at the workshop. The teams traffic-lighted the four frameworks (on the basis of the information provided in the case studies) against HEFCE's criteria of an adequate framework – which is to be credible and acceptable to the academic as well as user communities; to encompass the full range of economic, social, public policy, welfare, cultural and quality of life benefits, etc.

2. Case Study – Research Quality Framework

Context

- The Research Quality Framework was proposed by the Australian (Conservative) Government in 2004, with four aims:
 - Measuring quality of all taxpayer funded research;
 - Measuring the ‘impact’ of publicly funded research. This may include looking at the economic, social, environmental and cultural benefits of research in the wider community;
 - Measuring the quality of postgraduate research education and training;
 - Examining accessibility issues in relation to taxpayer-funded research, that is, issues relating to shared infrastructure, ICT, libraries and the accessibility of research results, etc.
- The RQF defined ‘impact’ as:
 - ‘The social, economic, environmental and/or cultural benefit of research to end users in the wider community regionally, nationally and/or internationally’.
- The RQF was abandoned in 2007 with a change in government and a manifesto commitment made by the incoming Labor administration
 - It is worth stressing that the RQF was abandoned for political reasons, not for feasibility reasons (although it was abandoned before an operational trial of the impact assessment model could be implemented).

In March 2005, Dr Brendan Nelson, the then Minister for Education, Science and Training, announced plans to develop the Research Quality and Accessibility Framework (RQF) to assess the quality and impact of research in Australia.² Nelson established a 13 member Expert Advisory Group (EAG) to advise on the development of the framework. The EAG published its preferred approach in September 2005 (Roberts et al., 2005).

Shortly after this report was released, Nelson was replaced as minister by Senator Julie Bishop. She responded to representations from a number of groups and announced the appointment of a new advisory group, the Development Advisory Group (DAG). The DAG established Impact and Metrics Working Groups to provide detailed advice on approaches to assess quality and impact. In October 2006, the DAG published its recommended approach for the RQF (Peacock et al., 2006a).

Development work continued throughout 2007 on the technical specifications for the framework. In the 2007 election campaign, the Australian Labor Party promised to remove the assessment of impact from any new funding model, and to move to a metrics-based approach. On winning power, they followed through with their election promises, and commenced development of a new initiative – Excellence in Research for Australia (ERA).

² It was “to provide a consistent and comprehensive approach to assessing publicly funded research” and would ultimately drive the distribution of research resources (Roberts et al., 2005).

Context

- In developing the process for the RQF, a Working Group on Research Impact (WGRI) was set up to advise the government on all issues relating to the assessment of impact.
- The WGRI produce a report that is no longer signposted or linked in the Australian Department of Education and Training (DEST) website, but is accessible by googling its title: 'Research Quality Framework: Assessing the quality and impact of research in Australia'.
- This short report is influenced by earlier work for the Australia National Health and Medical Research Council (NHMRC) in developing the MORIA (Model of Research Impact and Achievement) framework which also has not been fully implemented.

RAND

The DAG put in place a number of 'technical working groups' including the Technical Working Group on Research Impact (WGRI). The WGRI was established to devise an approach for assessing research impact, building on the EAG's proposals and recent developments in the field.³

The WGRI published its report which set out recommendations for assessing research impact (Peacock et al., 2006b). These are also summarised in the DAG report. Note that both reports are difficult to access on the web as they have been 'delinked' on the Department for Education, Science and Training (DEST) website as it is no longer official Australian Government policy.⁴

Because of the experimental nature of measuring impact on a national scale, it was felt essential to test the RQF impact assessment approach in an operational trial prior to its implementation. However, the RQF was abandoned due to a change of government in late 2007 before any official trial could be undertaken, although there were a number of university-led trials.

³ Group membership came from a wide range of backgrounds (in terms of disciplines and institutional affiliations), and included Professor Terry Nolan, from the University of Melbourne, who had led the National Health and Medical Research Council's Measurement of Research Impact and Achievement (MORIA) working group since its inception (Brutscher, Wooding and Grant, 2008). Much of the development work undertaken in MORIA informed and influenced the broader WGRI membership, as it was one of the few groups to have investigated the broad use of impact assessment and had reached the stage of proof-of-principle testing.

⁴ Under machinery of government changes the old DEST morphed into the DEEWR (Department of Education, Employment and Workplace Relations) with the change of government.

Principles

- 'The assessment of research impact is an incentive for researchers to focus on the outcomes of their research insofar as these provide benefit to Australia's society, economy, environment and/or culture'.
- 'The RQF seeks to identify and reward excellence in research, and is by definition inequitable. However, it is a fundamental principle that the impact assessment model will apply equally across all research fields and disciplines'.
- 'Impact assessment will complement research quality, as determined by the RQF, and while the underlying research must be demonstrably sound it does not necessarily have to score highly on Quality assessment to be rated for Impact'.
- 'Pure basic research will not be disadvantaged by the RQF, and neither shall this essential research be devalued. While Research Groupings focussed on pure basic research are unlikely to derive much benefit from the assessment of impact, they should be able to perform well in quality assessments. Conversely, Research Groupings that have an applied focus and may not achieve high ratings for quality, will have the opportunity to demonstrate their excellence through the impact assessment'.

The first section in the WGRI final report identifies eight 'underlying principles' for RQF to adhere to in assessing research impacts. These are repeated verbatim above and on the following slide and provide a good basis for developing an impact module for REF.

Principles (cont.)

- 'Research impact will be based on actual outcomes and their assessable impact. Prospective impact will not be part of the RQF'.
- 'Research which leads to tangible and assessable impact must, to some demonstrable degree, have been conducted recently, as the eventual impact assessment will be used to determine future research funding allocations'.
- 'A key principle is to direct RQF-related funding towards that research which generates impact. Hence, funding should be allocated towards institutions whose Research Groupings successfully demonstrate claims against any of the impact criteria. This will encourage behaviours that lead to a greater uptake of Australian research. Rewarding excellence is a further key principle, with enhanced funding distributed to institutions with Research Groupings receiving the highest impact ratings'.
- 'The natural unit of Impact assessment is the Research Grouping. These groupings already exist in many cases, and the grouping's rationale and strategy is an essential component of measuring Impact. It should be at the discretion of the research institution to select Research Groupings for the RQF, with the institution having discretion over how many Groupings and which staff are included for assessment. This entails that smaller Research Groupings should be accommodated, to a minimum of five group members with no upper limit'.

One of the key observations made in the principles is that the natural unit of assessment is a so-called 'Research Grouping'. Research Groupings would be at the discretion of the research institution to identify, including how many and which staff are included. It was envisaged that a Research Grouping should have no fewer than five members, with no upper limit.

It should be stressed that Research Groupings could be 'virtual' – that is, they might have little or no relationship to an academic unit. In fact, to meet the minimum of five staff in a discipline, smaller higher education institutions (HEIs) would almost certainly have to draw together staff from across faculties or departments. The larger HEIs had still not decided whether it was strategically better to have one large group consisting of maybe 50 or 100 staff, or whether it was better to split these down into smaller groups – certainly feasible in some disciplines. They were waiting to see how the funding would be allocated before making a final decision, but, as explained below, it never was.

What it is: a case study approach

- The WGRI concluded that 'impact metrics cannot be used as a proxy for determining impact ratings' and thus opted for a qualitative/case study approach. This conclusion is derived from a literature review by Donovan (2005).
- Each 'Research Grouping' would submit a Context Statement and an Impact Statement.
- The 'Context statement' for a Research Grouping is a word limited strategic case about the direction, focus, and nature of its research and how this relates to research impact.

RAND

The WGRI reviewed the literature and international practice (at the time, that is 2005/06) in assessing and demonstrating research impact (Donovan, 2005), and concluded impact metrics could *not* be used as a proxy for determining impact. It was particularly concerned about the use of metrics to identify the highest categories of impact. Thus, rather than prescribing a list of indicators to be used, the WGRI provided an indicative list of potential indicators that could help HEIs make their case for impact (these are listed on a following slide).

Instead of an indicator approach, it was accepted that the assessment of impact would rely on evidence-based impact statements, or case studies, containing both qualitative and quantitative information. It was noted that a Research Grouping's impact could not necessarily be assessed by using the same research outputs per researcher submitted as the basis for assessing research quality. So, the Expert Assessment Panels responsible for reviewing the case studies would appraise a Research Group's 'Context Statement' and 'Impact Statements'.

The structure of the Expert Assessment Panels would allow different, but overlapping membership for assessing research quality and research impact, with quality and impact judged on separate occasions. Each panel would have a core group of six members, three of whom would be international experts. Three additional discipline-specific experts would assist the core group for quality assessments, and another three expert end users would assist the core members in the assessment of impact.⁵

In its Context Statement, a Research Group would make a strategic case about the direction, focus and nature of its research, and how that relates to research impact.

⁵ The end users would come from industry, business, the public sector or community organisations. They would provide a broad perspective on the determination of ratings for quality and impact and must be separate from immediate collaborators of a Research Group.

What it is: the 'Impact Statement'

- The 'Impact Statement' addresses four key questions:
 - How has the Research Grouping engaged with end users to address a social, economic, environmental and/or cultural issues?
 - What new products, policies, legislation, paradigm, attitudes, outlooks, etc have end users adopted, implemented and/or taken up as a result of engaging with your Research Groups?
 - What are the social, economic, environmental and/or cultural benefits of the new product, policy, legislation, paradigm, attitude, outlook, etc adopted by end users?
 - What is the magnitude or extent of social, economic, environmental and/or cultural benefit to end users as a result of the implemented research?
- In addition to addressing the four questions, each Research Grouping provides up to four case studies that illustrate and support the claimed and a list of end users that could be contacted by the Expert Assessment Panels.
- The WGRI provided a list of potential impact indicators which they stress are illustrative and non exhaustive but could be used to support claims of impact.

The basis for the assessment of impact would be an Impact Statement of up to ten pages which includes verifiable, evidence-based claims against specific impact criteria (listed in the slide overleaf), for up to four case studies. The case studies were meant as illustrative examples of the claims made.

It was specified that research impact would be assessed on actual outcomes and their assessable impact, as opposed to potential impacts. The impact to be assessed would have to have occurred within the six-year assessment period, but could be based on original research that was conducted six years prior to the assessment period.

It was anticipated that Research Groupings would use the list of indicative impact indicators (slide below) to help provide verifiable evidence that the impact occurred and to demonstrate the link between the claimed impact and the Group's research. HEIs were to provide details of end users of the research who could be approached for the sample of case studies that it was proposed to audit. HEIs were explicitly told not to approach end users to provide statements of support prior to submission. While the business community was supportive of the process, the government did not wish to alienate them with undue burdens on their time – much of which would, in any case, be unnecessary, as other supporting information would be more than sufficient to support the HEI's claim.

What it is: potential indicators

- Reduced pollution
- Regeneration or arrested degradation of natural resources
- Lives saved
- Reduced infection rates
- Reduced treatment time and cost
- Increased literacy and numeracy rates
- Positive reviews of creative publications and performances
- Increased cultural awareness
- Royalties
- Increased employment
- Reduced costs and resource usage
- Increased competitiveness of Australian industry
- Spin off companies
- New products & inventions
- Licences
- Creative works commissioned
- Change in procedures, behaviours, outlook, etc
- New policies, guidelines, legislation etc
- Citations of research in legal judgements which become case law
- Contracts and industry funding
- Repeat business
- Number of presentations involving contact with end users
- Community awareness of research
- Non-academic publications & performances
- Collaborative projects with end users
- Citations in Government reports, Hansard, etc
- Provision of expert advice and submissions to enquiries, etc
- Invitations to be a visiting researcher or researcher in residence at an end user institution (if based on research merit)
- Invitations to be on reference, advisory and/or steering committees (if based on research merit).

RAND

It was felt that defining a list of acceptable indicators was likely to disenfranchise many disciplines, or even whole fields. For example, selecting indicators with a focus on spin-off companies, product development and so on is likely to overlook virtually all research in the social sciences and humanities. It was also felt that it could result in an unnecessary burden on HEIs to collect information that was totally irrelevant for many of their impact studies.

Where it is being used

- In preparing for the RQF, three universities – Griffith, Flanders and James Cook University – ran a pilot.
- This included preparing material on how to develop ‘best practice’ case studies, and having others review the case studies.
- Opportunity for HEFCE to learn from these pilots, especially given that there was no operational pilot as planned, given the abandonment of the initiative.

RAND

The Innovative Research Universities of Australia (IRUA) group of HEIs actively supported the assessment of impact alongside quality in the RQF. They undertook extensive trialling of impact statements, with a view to producing a template of ‘best practice’ for writing context statements and demonstrating the link between the research and the claimed impact. The trials were conducted, but the RQF was abandoned before any final report was written. There are a number of people within the HEIs involved who can provide feedback about the process, particularly at Griffith and James Cook universities.

How it is being used

- A key principal was to direct funding towards research which generates impact.
- Related to this was an acknowledgement that the funding would not be linear – the high impact institutions would receive proportionally more funding.
- No specific amount had been set in terms of the weighting for the impact assessment element of the final funding formula, although it was expected to be around 10 percent possibly increasing as the methodology was developed.

RAND

Detailed technical work on funding was to be continued through early 2007, but was never published due to the change in government. Although no specific amount had been determined for the weighting of the impact assessment element of a funding formula, it was expected to be around 10 percent possibly increasing as the approach was developed. The DAG, however, agreed on the following principles to determine funding from the RQF outcomes (ie including quality as well as impact):

1. Funding will be allocated using a relative funding model similar to that of the current research block grants, but based on assessed Research Groups rather than at the level of the total Institution.
2. RQF quality and impact ratings for Research Groups will be separately aggregated to the institutional level for RQF funding purposes.
3. In order to reward high quality and impact research, funding will be distributed with higher weighting for higher RQF ratings in a linear fashion (with the precise gradient to be determined).
4. Research groups rated below '2' for quality will not contribute to an institution's RQF quality funding.
5. Research groups rated below 'D' for impact will not contribute to an institution's RQF impact funding. (Please see next slide.)
6. Funding will take into account the cost of the research assessed, possibly based on either Research Assessment Exercise or Performance Based Research Fund discipline weightings.
7. Funding will take into account the volume of an institution's research, as measured by its full-time equivalent staff assessed for the RQF.
8. HEIs will retain discretion over the internal allocation of RQF-driven Institutional Grants Scheme and Research Training Scheme block funding.

Assessment output

- Expert Assessment Panels would review the Research Groupings' Evidence Portfolios to make an assessment on research quality & impact.
- Based on the evidence provided in the Context and Impact Statements the Expert Assessment Panel would determine the validity of the claims made against impact ratings (see below).
- Each panel would have six core members, three of whom were international plus three who are discipline experts (and will assess quality) three who are end users to assess impact.

Rating	Description
A	Adoption of the research has produced outstanding (in the top 2%) identifiable social, economic, environmental and/or cultural benefit for the wider community regionally, nationally or internationally.
B	Adoption of the research has produced identifiable social, economic, environmental and/or cultural benefit for the wider community regionally, nationally or internationally.
C	Research has been adopted to produce new policies, products, paradigms, attitudes, behaviours and/or outlooks in the end-user community.
D	Research is engaging with the end-user community to address social, economic, environmental and/or cultural issues regionally, nationally or internationally.
E	Research has had limited or no identifiable social, economic, environmental and/or cultural outcome for the wider community regionally, nationally or internationally.

RAND

The Expert Assessment Panels were meant to assess the broad impact of research on the basis of the five-point scale provide in the slide above. The panels would also be able to call on the advice of external experts if they felt unqualified to assess some of the work before them. This was likely to be quite common, as the panels had to cover a broad sweep of disciplines: for example, one panel had to cover all of physics, chemistry and earth sciences. DEST had been in the process of establishing a list of experts, primarily drawn from the Australian Research Council and National Health and Medical Research Council Australia (NHMRC) assessors, when the RQF was abandoned.

It should be noted that there was some debate on the WGRI about the rating scale illustrated above. This had been drawn up by the DAG and had ministerial approval prior to the working group meeting so it could not be modified, though it was the subject of much discussion.

Stakeholders involvement

- Stakeholders are heavily involved in the RQF:
 - Evaluatees provide the case study examples of impact and evidence.
 - End users form part of the Expert Assessment Panels.
 - They also act as referees for Research Groupings.
- There was a strong believe within WGRI and DEST, that a methodology could be developed. Feedback on the trials has never been published.

RAND

Stakeholders would be heavily involved in the RQF – with evaluatees submitting Context and Impact Statements and end users sitting on the Expert Assessment Panels (together with international experts) and acting as referees for Research Groupings.

Within the WGRI and DEST, there was a belief that an approach could be developed – without down-playing the difficulties of the task. However, it was also acknowledged that it would need to be an assessment of a sample of impact outcomes, not a measurement of the total impact achieved. The latter was clearly impossible because of the case study approach required, but as such represented a disparity with the methodology to be employed for quality assessment, which, through the use of metrics, did seek to assess all output.

Additional technical aspects

- The programmes are assessed and reassessed on a six-year schedule.
- Attribution: the RQF addresses the issue by means of a sampling approach.
- Time frame: impacts have to occur within the six-year assessment period, but can be based on research up to six years prior to assessment period.
- How many submissions: a Research Grouping can submit four case studies at a time.
- Additionality: not addressed.
- Scalability: RQF would have applied for Australian HEI system.
- Application across areas: it was intended to apply across all research areas.
- Decision making: funding decisions (c 10 percent) were meant to be influenced by impact assessment.
- Auditing: Research Groupings have to provide contact details of end users that can act as referees.
- Burden: not known.

RAND

One of the biggest issues in relation to the assessment of impact is that of attribution. This has two elements: the problem of proving the claimed impact resulted from a particular piece of research and deciding which HEI can claim the research on which the impact is based.

In some instances, the link between the original research and the demonstrable benefits are clear. But this is the exception rather than the rule. More often than not, impact is achieved through the implementation of the results of a number of research projects from a diversity of HEIs over a long time frame. This is the primary reason why the assessment of impact is usually put into the 'too hard' basket.

The RQF approach for dealing with this issue was to make explicit that the assessment was using a sampling approach. It did not attempt to measure all impact, but rather sought to demonstrate the types of benefits that university research had achieved. It allowed HEIs to select a relatively small number of research outcomes where the path through from research to impact was more readily and verifiably demonstrable.

As noted earlier, in terms of the second type of attribution 'problem' the WGRI came up with guidance on which impacts HEIs could claim over what time period: Impacts have to occur within the six-year assessment period, but can be based on research up to six years prior to the assessment.

In terms of auditing, Research Groupings were supposed to submit with their Impact Statement details of end users who can be contacted as referees.

Observations/lessons learnt

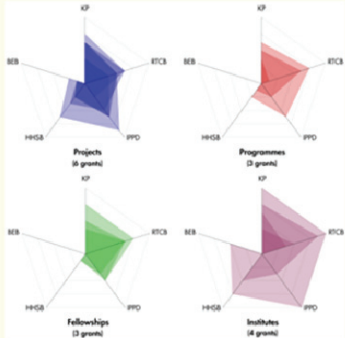
- The RQF was abandoned for political not feasibility reasons.
- The RQF process was 'worked up' and part piloted when it was abandoned.
- There is some excellent thinking that can be applied to the REF.
- Access to documentation is difficult (either via contacts, semi random Google searchers, existing personal literature, or 'de linked' DEST website)
 - Can HEFCE request full documentation from DEST?
- Some learning – especially on the pilot – will be 'under the skin'.
 - May be worth investment for HEFCE to spend 'time on the ground'
- Key lessons are: use of case studies, rejection of 'impact indicators' as a stand alone methodology, but provision of potential indicators that can be used to support claims of impact .

The RQF provides an obvious basis to develop an impact module for the REF. Much work was undertaken before it was abandoned for political reasons. Some of this work is inaccessible and HEFCE may benefit from direct discussions with DEEWR to see if it would be possible to transfer some of the learning. Likewise, there was a lot of work undertaken in HEIs that could inform the development of an impact module for REF. Again, HEFCE may benefit from direct discussions with those HEIs.

3. Case Study – RAND/ARC Impact Scoring System

Context

- Arthritis Research Campaign (ARC) wanted to improve end of grant reporting and track impacts of their research over time:
 - ARC is the fourth largest biomedical charity funder in UK, over £25m pa
 - Funds basic research, clinical research and research by allied health professionals
- Development started in 2006 and new system was implemented in 2008.
- Follow on from case study based project to develop detailed understanding of impact of ARC funding.



Payback Profile

RAND

The Arthritis Research Campaign (ARC) is a medical research charity in the UK with the mission of curing arthritis. The RAND/ARC Impact Scoring System (RAISS) is used to collect information on the outputs and impacts of ARC grants. Development by RAND Europe and ARC started in 2006 and the final approach was introduced in 2008 to replace ARC's end-of-grant reporting system. The approach places particular emphasis on wider impacts such as improvements in research capacity, effects of policy and health interventions. The development of the approach is summarised by Wooding (2008) and described in detail in a forthcoming report (Wooding et al., forthcoming).

The project to develop a tool to map ARC grant outcomes followed on from an earlier project that sought to understand the impacts of a small group of historic ARC grants through case study research (Wooding et al., 2004; 2005). That project summarised its findings through 'Payback Profiles' (as illustrated in the slide above) that aimed to quantify the level of impact produced from each grant in five categories: Knowledge Production (KP), Research Targeting and Capacity Building (RTCB), Informing Policy and Product Development (IPPD), Health and Health Sector Benefit (HHSB) and Broader Economic Benefit (BEB). The profiles shown on the slide group the 16 grants by funding type and overlay the payback profiles of individual grants of the same funding type.

Principles

- Should be applicable to all areas of ARC research
- Should be applied to every grant
- Should collect only what can be analysed
- Should minimise the burden on researchers and administration resource
- Should provide an overview of ARC portfolio to:
 - Inform strategy
 - Influence future decisions about funding
 - Form foundation for detailed evaluation.

RAND

In developing the survey instrument, RAND Europe worked closely with a number of active ARC researchers and piloted versions of the instrument at various stages in the process. Throughout the development process we were guided by a set of core principles agreed with ARC.

Of particular note was that the ARC was keen that the approach should be applicable to all areas of ARC research. Although restricted to the medical field, ARC research covers a very broad spectrum of both areas of biology and stages of research, from basic laboratory research, through clinical research to research carried out by allied health professionals such as physiotherapists.

Learning from the experience of end-of-grant reports that collected large amounts of information but in a relatively unstructured way, and hence were very hard to analyse, it was decided that the tool should collect only information that could be analysed and should try to minimise the burden of both collecting and analysing this data.

It was intended that the information from the tool would be used at the aggregate level to provide an overview of ARC's portfolio to inform strategy, influence future decisions about funding at a discipline or area level and provide a firm foundation for more detailed evaluations. It was not envisaged that information from the tool would be used to evaluate specific grants, or feed into specific funding decisions.

What it is

- A web survey based impact questionnaire aiming to capture overview information on grants awarded by the Arthritis Research Campaign:
 - 187 Yes/No questions covering scientific, medical and health impacts. No description of impact is collected
 - Captures only impacts that have occurred, does not capture process
 - Depends entirely on honesty and accuracy of grant holders.
- Aims to show ARC the range and distribution of impacts to inform strategic decision making.

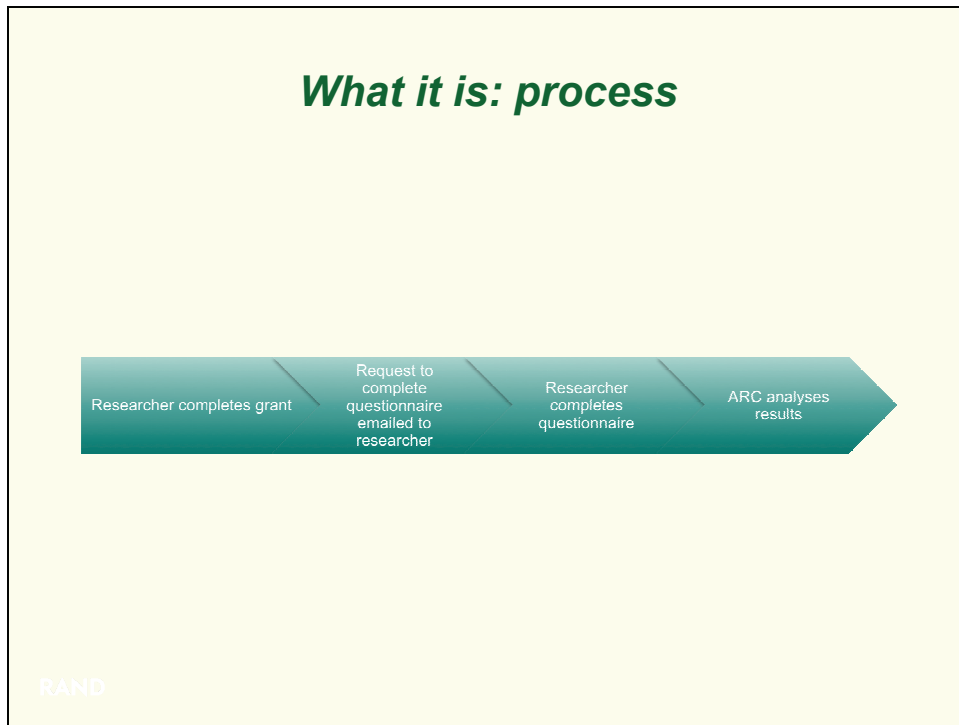
RAND

The impact mapping portion of the approach consists of 187 Yes/No questions that aim to capture a broad range of possible outcomes that have arisen from each research grant. The questions cover:

- what further research has developed from the grant
- what research tools have been developed
- what dissemination activities have been carried out
- what impacts on health policy have occurred
- how the grant has affected education; and what health interventions
- changes in service delivery or health advice produced.

Over 90 percent of investigators can complete the impact cataloguing tool in less than 60 minutes. The web-based questionnaire is completed by the principal investigator of the grant and relies entirely on this source of information. It asks principally about impacts that have occurred, rather than potential or future impacts, and it captures only the existence of an impact, it does not collect any descriptive information beyond that in the question.

By analysing the data across the portfolio of research ARC funds, the tool allows ARC to understand the distribution and range of impacts arising from its research to inform strategic decision making.



The administration of the approach is very simple. Six months after a principal investigator completes a grant, ARC sends them a request to complete the questionnaire by email. This email contains a direct link to a personalised website where the researcher can complete the questionnaire for the grant. These results are fed directly into a database from which ARC can then extract the data to examine the impacts of its research.

What it is: example questions

5.4: Dissemination - Other Dissemination

98. Dissemination to health professionals?
 Yes No Not Known

If yes, has this included (If no, go to question 5):

99. Websites for health professionals?
 Yes No Not Known

100. Printed material (such as booklets) for health professionals?
 Yes No Not Known

101. If yes, what was the distribution of this printed material? If no, go to question 99:

	Yes
Local	<input type="radio"/>
National	<input type="radio"/>

102. Dissemination to patients/public?
 Yes No Not Known

If yes, has this included (If no, question 9):

103. Websites?
 Yes No Not Known

104. Printed material (such as booklets) for patients/public?

Page 40

8.1: Interventions/Products - IP

140. For the research cited by patents, or intellectual property protection, which stages were carried out during the grant, or have happened since?

	Yes	No	Not Known
Cited in a patent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Patent applied for/filed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Patent granted	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Patent licensed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Page 41

8.2: Interventions/Products - Pharmaceutical

141. For the therapeutic pharmaceutical product, which stages were carried out during the grant, or have happened since?

	Yes	No	Not Known
animal tests	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
phase 1 clinical trial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
phase 2 clinical trial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
phase 3 clinical trial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
on the market	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
trials a current pharmaceutical for a new indication or with a new regime	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

This slide provides a series of example questions from the survey to illustrate how various different areas of impact are covered. A full set of questions is available in the report describing the development of the tool (Wooding et al., forthcoming).

Where it is being used

- Who developed it? RAND Europe and ARC
- Who is using it?
 - ARC is using it as an end of grant reporting system
 - MRC is using a similarly structured tool that also collects more descriptive information to collate its past impacts.
- It is a relatively new tool, but there is interest from other biomedical research funders.

RAND

In addition to its use by ARC as an end-of-grant reporting tool, the Medical Research Council (MRC) built on the structure of the tool to develop their Outputs Data Gathering Tool (ODGT), which they are now using to collect information on the wider impact of MRC grants. The MRC added additional questions that collect qualitative descriptions of the impacts. More details on the ODGT is available from the MRC website (Medical Research Council, 2008).

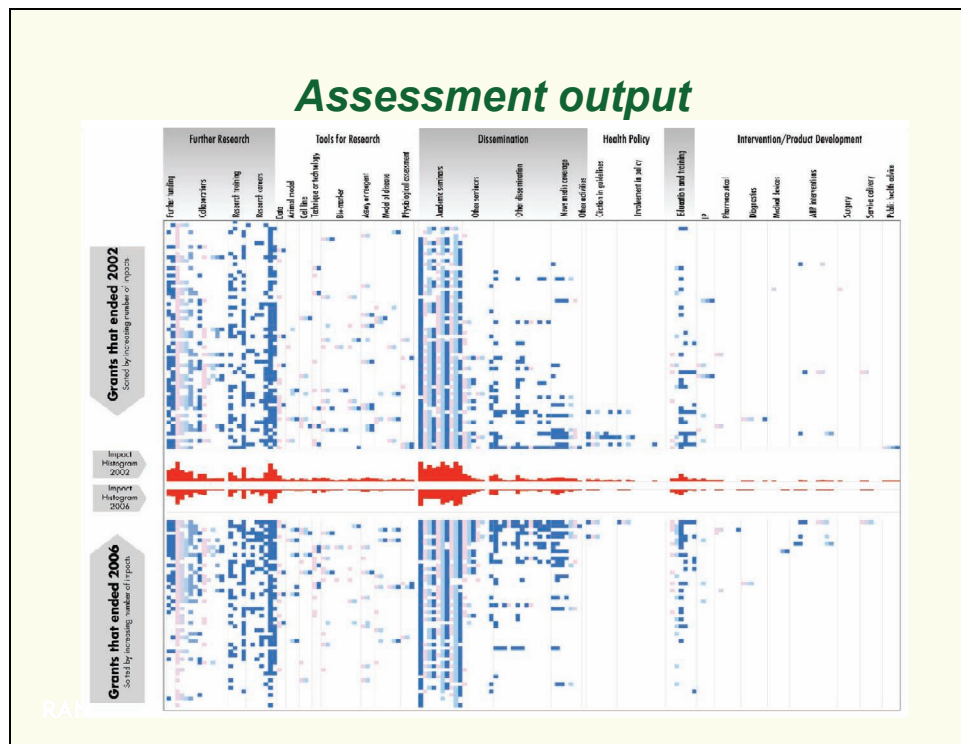
In addition, there has been interest from further research funders in using the tool or adapting it to cover other areas of research.

How it is being used

- Consequences:
 - Assessment serves as input to strategic decision making
 - There is no direct link to funding, however.

RAND

RAISS is designed to feed into strategic decision making and provide a basis for more detailed evaluation (for example, by providing information for case study selection); however, it is not designed to provide input to specific grant-making decisions. As there is no direct link to funding the dependence on information from the principal investigators is seen as an acceptable trade-off.



The tool was initially piloted by applying it retrospectively to all 136 ARC grants that finished in 2002 and 2006. The completion rate for the survey was 87 percent, producing 118 responses; these results can be visualised as the Impact Array shown above. The Impact Array illustrates the range of impacts generated by each grant. The assessed impacts include: levels of collaboration, additional funding, career progression, publications, citations, policy outcomes and generation of tools for future research (see Impact categories along the top of the array).

Each impact is represented by a coloured block in the graphic – from pink to blue, with pink representing early stages or local effect, and dark blue being developed or international impact.

Each grant is represented by a single row of pink to blue blocks, one for each type of impact. The grants are sorted by total number of impacts, as indicated by the arrows, and so the chart lets you identify those with the highest number of impacts (the grants details are not shown here, just their impacts).

This grants that finished in 2002 and in 2006 have been grouped separately to allow comparison. The red histogram along the central part of the chart shows the total number of each type of impacts for all the grants in each section and allows us to compare 2002 and 2006 impacts.

Stakeholder involvement

- Stakeholders:
 - System is run by the grant funder (ARC)
 - Impacts defined by funder
 - Completed by the grant holder
 - No other stakeholders, eg users, are involved
 - Researchers seem positive (few negative comments), funder seems positive.

RAND

RAISS is very much structured as a way for a funder to collect data on the impacts of their research from the researchers who carried out that research. Other stakeholders are not involved in the approach and their list of types of impacts is detailed and pre-defined although there is the option for researchers to suggest additional categories that have been omitted. Reaction to the tool by researchers has been generally positive, with fewer than ten negative comments received.

Additional technical aspects

- In use as a replacement for end of grant reporting, two waves of retrospective evaluation on grants ending in 2002 and 2006.
- Attribution: cut-off at 'significant contribution' level. **Hard to include.**
- Time frame: arbitrary, relies on recall; used for 2 years and 5 years. **Flexible.**
- How many submissions: one for each grant, 118 in retrospective pilot. **Could use at other levels of aggregation.**
- Additionality: not addressed. **Hard to include.**
- Scalability: easily scalable.
- Application across areas: only developed for biomedical science. **Challenging to extend across all disciplines.**
- Decision making: currently provides distribution of impacts, does not value impacts. **Would require addition of peer review.**
- Auditing: Currently, none. **Could be audited by sample.**
- Burden: low – 30–60 minutes per submission.

RAND

The slide above outlines a variety of technical aspects with regard to the operation and implementation of the approach. In black we outline how the approach is currently being used with the pink text summarising the issues in extending the approach to a nation wide research impact evaluation approach. We discuss each issue in more detail below.

The tool was initially piloted to capture the impacts of all ARC grants that finished in 2002 and 2006 and it is now in use to capture the impacts of all grants six months after their completion.

Attribution

The tool has a relatively simplistic approach to attribution asking for impacts to be mentioned where the grant in question has made a 'significant' contribution to the impact.

Time frame

The tool could be applied to grants that had finished at any point, the only limitation being availability of the principal investigator and the quality of their recall.

How many submissions

Currently the approach is used to catalogue impacts at the level of each individual grant – this means there will be double counting where one impact has been contributed to by a number of grants and under counting where one grant has lead to more than one impact of a particular type.

Additionality

The tool does not attempt to address the issue of additionality and it would be hard to include.

Scalability

The tool is easily scalable to a larger number of grants; it is simply an issue of inviting the additional researchers to complete the survey.

Application across areas

Building a comprehensive impact list for arthritis research was challenging and there were areas – such as actual health benefits – which were abandoned, as it was not possible to devise suitable Yes/No questions. Devising a set of questions that encompassed all disciplines would be a difficult task and risk confronting the researchers with an unmanageably large list of possible impacts to review.⁶

Decision making

Currently the tool provides a mapping of impacts but does not attempt to put a relative value on the different types and kinds of impacts. Hence it is impossible to ‘aggregate up’ to give a meaningful total impact score for each grant. This limited the ability of the process to feed into Yes/No or formula-based decision making.

Auditing

The approach is not currently audited, although there is no reason why it could not be audited through follow up of a sample of responses.

Burden

The approach is quick for the researchers to complete, with over 90 percent of submissions taking less than an hour.

⁶ The unmanageably large list would surely be overcome by having separate lists for each discipline; though the huge task of devising suitable ones across all disciplines still remains.

4. Case Study – Program Assessment Rating Tool

Context

- PART can be viewed as evolving out of the Government Performance Results Act (GPRA) of 1993, which mandated the development of a system for assessing performance of all government programmes.
- PART was developed and is run by the Office of Management and Budget. Its development started in 2002. It was first introduced in 2004.
- In the first budget the effectiveness of 234 major federal programmes was numerically rated (which corresponds to about 20 percent).
- In each of the succeeding budgets, approximately 200 additional programmes were assessed, for a total of about 1000 as of 2008.

RAND

The Program Assessment Rating Tool (PART) in the US can be viewed as evolving out of the Government Performance Results Act (GPRA) of 1993. GPRA mandated the development of an approach for assessing performance of all government programmes.

Shortly after George W. Bush took office in 2001, he committed to an agenda of improved government management. A key element of this agenda was to make government more results-oriented by expanding the use of performance budgeting (ie allocating funding on the basis of performance). In a first step, Bush directed the Office of Management and Budget (OMB) to work with each agency to recast its budget to include performance information.

In 2003, he expanded this effort by committing to a programme-by-programme assessment of performance. Again, he directed the OMB to take a lead on this assessment effort. In response, the OMB developed an assessment framework, with the assistance of agencies and outside experts, named the Program Assessment Rating Tool, or PART.

PART is now in its sixth year and is used also but not only to assess the performance of federal research and development programmes. In the first year the effectiveness of 234 major federal programmes was assessed. This equalled about 20 percent of all federal programmes at the time. In each of the succeeding budgets, approximately 200 additional programmes were assessed. Today over 1000 programmes have been assessed by PART. Programmes are assessed at least once every five years.

Principles

- A PART review helps identify a programme's **strengths and weaknesses** to **inform funding** and **management decisions** aimed at making the programme more effective.
- It looks at **all factors that affect and reflect a programme's performance** including its purpose and design; performance measurement, evaluations, and strategic planning; programme management; and programme results and accountability.
- The PART is designed to provide a consistent approach to assessing and rating programmes **across the Federal government**.
- PART's focus is on **outcomes/impacts** and **efficiency**.

RAND

The key principles/ambitions of PART are:

- to improve government management
- to induce organisational change
- to be as holistic as possible
- to be as general as possible
- to be results/outcome driven.

As mentioned above, the PART framework was developed primarily to improve government management. The idea was to use PART to provide decision makers with the information they need to better allocate scarce resources: 'Even modest success in identifying programs that are effective and those that are ineffective, and facilitating some movement of money away from the ineffective and towards the effective, will be a valuable accomplishment' (Gilmour, 2006).

A second principle/ambition (no less relevant to HEFCE) was to induce organisational change – that is to encourage agencies/organisations to find better ways of achieving their goals and to improve their results. One way to achieve this is by using PART to inform funding decisions. If funding decisions are based on PART's assessment of a programme's performance, this provides a strong incentive for programme officials to improve their programme's performance – and introduce change.

Another way that PART contributes to organisational change is by providing programmes not only with a numerical score after an assessment, but also with an 'improvement plan' – that is, a set of management, policy, budget or legislative follow-up actions created to respond to assessment findings by the OMB. An improvement plan contains as many as three follow-up actions.

The third principle/ambition of PART is to be as holistic as possible. As will be explained in the next slide, PART does not only assess outputs, outcomes and impacts, but also a programme's purpose, objectives and management. This is a key difference to the three other frameworks reviewed in this report. A positive aspect of this is that it reduces the danger of imposing set purpose and goals upon programmes (or research groupings in HEFCE's case) which do not get any buy-in from officials and/or steering programmes (groupings) in a not-fruitful direction. The latter point is particularly dangerous if central government (ie the Office of Management and Budget) lacks relevant expertise.

PART also aims to be general. It is applied in the same way across all federal government, although sometimes small adjustments will be made. PART has been applied to a range of programmes, including the following:

- competitive grant programmes
- block/formula grant programmes
- regulatory-based programmes
- capital assets and service acquisition programmes
- credit programmes
- research and development programmes
- direct federal programmes.

Finally, PART aims to capture outcomes/impacts rather than outputs (which makes it a relevant case study for HEFCE). The Office of Management and Budget clearly states its preference of outcome/impact measures over output measures: 'Outcome [/Impact] measures are most informative, because these are the ultimate result of a program that benefits the public. Programs must try to translate existing measures that focus on outputs into outcome [/impact] measures by focusing on the ultimate goal of the program.' It defines outcomes/impacts as: '[...] the intended result of carrying out a program or activity. They define an event or condition that is external to the program or activity and that is of direct importance to the intended beneficiaries and/or the public', (OMB, 2005).

What it is: the questionnaire

- PART takes the form of a questionnaire.
- The questionnaire contains 25–30 general (Yes/No/NA) questions about each of the following topics:
 - Programme purpose and design
Is the programme purpose clear?
Does the programme address a problem, interest or need?
 - Strategic planning
Does the programme have specific long-term performance measures?
Does the programme have ambitious targets and timeframes?
 - Programme management
Does the agency collect timely and credible performance information?
Does the programme collaborate with related programmes?
 - Programme results
Has the programme demonstrated progress in achieving its long-term performance goals?
Does the programme achieve its annual performance goals?

RAND

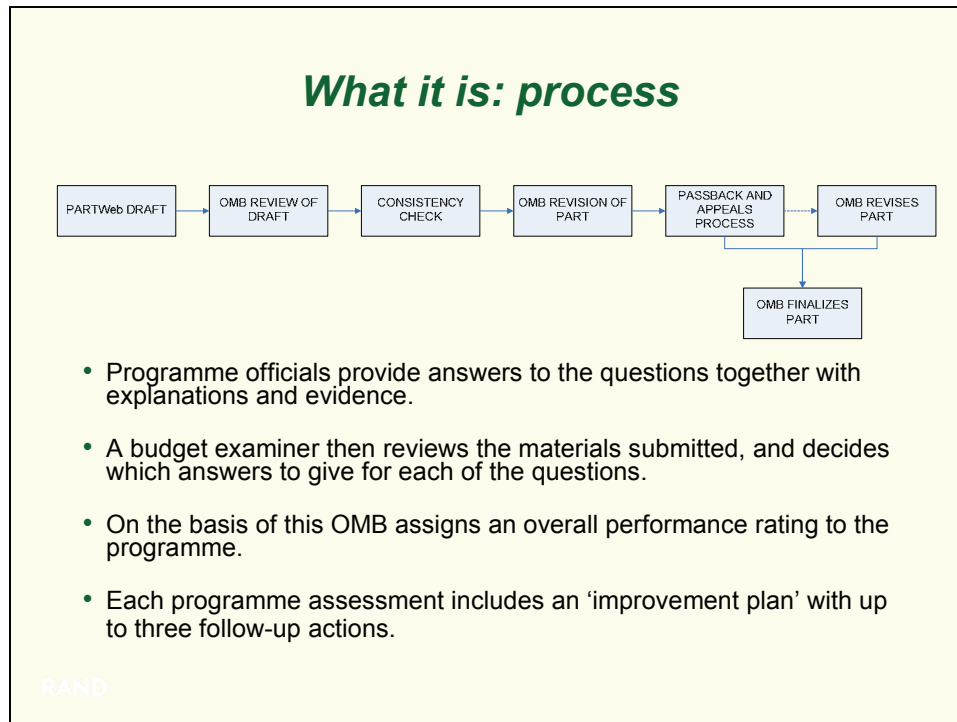
PART is an example for a systematised self-evaluation approach to assessing impact. It takes the form of a diagnostic questionnaire (OMB, 2005). The questionnaire contains 25–30 general questions about each of the following four broad topics:

- Programme purpose and design (20 percent): to assess whether the programme design and purpose are clear and defensible. (Sample question: Does the programme address a specific and existing problem, interest or need?)
- Strategic planning (10 percent): to assess whether the agency sets valid annual milestones and long-term goals for the programme. (Sample question: Is the programme designed so that it is not redundant or duplicative of any other federal, state, local or private effort?)
- Programme management (20 percent): to rate agency management of the programme, including financial oversight and programme improvement efforts. (Sample question: Does the programme use strong financial management practices?)
- Programme results (50 percent): to rate programme performance on goals reviewed in the strategic planning section and through other evaluations. (Sample question: Has the programme demonstrated adequate progress in achieving its long-term performance goals?).

The questions in sections one to three can be answered with Yes, No or Not Applicable (where the latter has to be justified). Section four allows for answers: Yes, Large Extent, Small Extent, No.

Each section carries a pre-specified weight (see percentages in brackets above) resulting in a total weighted numerical rating ranging from 0 to 100. Programme managers can alter weights within each category to emphasize key factors of the programme. To avoid manipulation of the total score, weights must be adjusted prior to responding to any question.

Based upon the numerical scores, OMB assigns a management and performance rating to each programme (again, including research and development programmes). These range from the highest rating of 'effective', to 'moderately effective', to 'adequate', to a lowest score of 'ineffective'. In addition, the rating of 'results not demonstrated' means that the measures developed were not adequate to determine the programme's effectiveness.



The process of PART is relatively straightforward: programme officials suggest answers to each question in the questionnaire and provide explanations and evidence supporting their answers. A budget examiner for the programme then reviews the materials submitted, and decides, whether or not it warrants the suggested answer and, which answers to give for each of the questions. The figure in the slide above gives a graphical representation of the process.

Federal agencies have the possibility of formally appealing the answers with which they disagree. Appeals are considered and adjudicated by a five-person panel comprised of members of the President's Management Council, a group of deputy secretaries responsible for management issues at their respective agencies.

Where it is being used

- Who developed it?
 - The Office of Management and Budget.
- Who is using it?
 - The Office of Management and Budget
 - The Administration provides PART assessment details also to the public and Congress
 - In general, because the PART is an evidence-based analysis, the Administration believes it can be a useful tool in budget and policy formulation for all branches of government.
- There have been only smaller changes to PART over time. It is still unclear what will happen to PART under the new administration.

RAND

PART is run, and its results used by, the Office of Management and Budget. There is a notion that, because PART is an evidence-based tool, it can be useful in budget and policy formulation beyond the OMB. In recent years there have been attempts in particular to get buy-in from Congress, but, so far with mixed outcome (Kniker, 2006).

There have been only small changes to PART over time. It is not clear whether PART will be continued under the Obama administration.

How it is being used

- Consequences:
 - Assessment serves as an input to strategic decision making
 - There is also a link to funding decisions. It is, however, not deterministic.

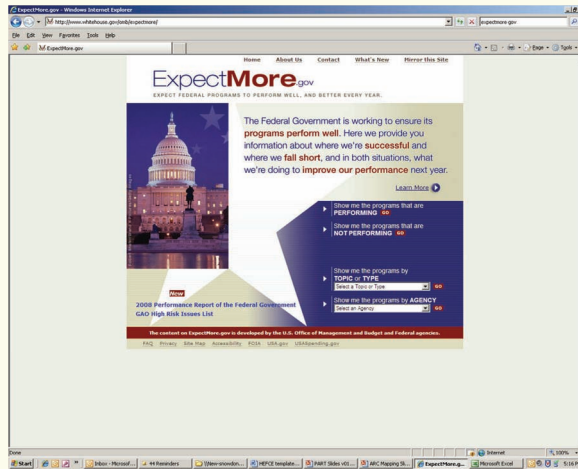
RAND

One of the key principles of PART is to inform funding decisions. It is important to note, however, that, because there are a number of factors contributing to a programme's budget request, a good PART rating does not automatically guarantee a specific level of funding:

'A program may be Effective, but if it has completed its mission, if it is unnecessarily duplicative of other programs, or if there are higher priorities, its funding may be reduced. Likewise, an Ineffective or Results Not Demonstrated (RND) rating does not guarantee decreased funding. For example, an RND-rated program may receive additional funding to address its deficiencies and improve its performance (*Expect More*: www.expectmore.gov).

Assessment output

- The website ExpectMore.gov is the primary public mechanism for reporting on Federal programme performance and what is being done to improve results.
- The site includes simple mechanisms that enable the user to identify programmes that are performing or not performing.



RAND

In February 2006, the OMB unveiled a new website, www.ExpectMore.gov. It makes available the assessments of all programmes that have been subjected to PART. The website is the primary public mechanism for reporting on federal programme performance and what is being done to improve results.

The site includes a variety of simple mechanisms that enable the user to identify programmes that are performing or not performing. The idea is to positively acknowledge programmes that are performing and – for example, by giving stakeholders arguments to demand improvement – to put pressure to improve on programmes that are not (Trochim, 2006).

ExpectMore.gov is designed to be a user-friendly public site, which provides simple explanations for the ratings. It uses different numbers of stars to indicate the final rating category for each programme. The user can drill down on a particular overall rating to get more detail about how that programme's assessment was reached and what the improvement plan is.

It is not difficult to see how such a mechanism could be used to report REF assessments of research groupings.

The next slide provides an example of the PART rating details for a programme, which was judged to be ineffective – the 'Even Start' programme.

Assessment outputs (cont.)

- PART provides an overall score.
- The users can drill down to get more detail about how that assessment was reached and what the improvement plan is for the programme.



RAND

For each assessment a summary is provided on ExpectMore.gov. It provides a programme overview, some of the key findings of the assessment, and the follow-up actions agencies are taking to address those findings. In addition, each summary links to the full programme assessment details, the programme's improvement plan, the programme's website (where available) and the search results for similar federal programmes.

Stakeholder involvement

- Stakeholders are heavily involved in PART:
 - Evaluatees provide the definition of programmes, strategic goals and annual targets.
 - They also provide (evaluation) evidence for these goals and targets.
 - In addition, evaluatees have the possibility to appeal to the assessment.
 - Gilmour (2006) finds: ‘the analysts and program managers interviewed by the author [...] almost uniformly believed that the exercise of completing the PART questionnaire was good for programs’.

RAND

Evaluatees are heavily involved in the PART process: They provide suggested answers to the questions in the questionnaire and evidence to support them. In addition, evaluatees have the possibility to appeal an assessment. If they can demonstrate significant improvement, they can also request a reassessment to improve the rating of their programme after one year.

Trochim (2006) finds that ‘[as] would be expected with any federal program of such magnitude, the PART process and the ExpectMore.gov website have not been without controversy. Several reports of the Government Accountability Office have raised concerns about implementation and advocacy groups outside the federal government, such as OMB Watch, remain critical’, (Trochim, 2006).

Overall, however, as Gilmour (2006) points out, even though ‘management systems imposed from above always meet a certain amount of scepticism and resistance [...] attitudes have changed as program managers have seen the determination and persistence of OMB in implementing PART. [...] the analysts and program managers interviewed by the author – virtually all careerists – almost uniformly believed that the exercise of completing the PART questionnaire was good for programs’ (Gilmour, 2006).

Additional technical aspects

- The programmes are assessed and reassessed on a five-year schedule.
- Attribution: it is up to the programmes to use an appropriate evaluation methodology (to get credited).
- Time frame: to the extent that what counts is whether programmes meet their annual (outcome) targets, it does not matter how far back the drivers for changes go.
- How many submissions: a programme is supposed to have a limited number of long-term goals (and needs to demonstrate effectiveness against each).
- Additionality: input and output/outcome additionality addressed. Behavioural additionality is not an issue.
- Scalability: PART is used across government.
- Decision making: funding decisions are not solely based on PART, but PART enters decision making process.
- Auditing: up to discretion of OMB examiner responsible for programme.
- Burden: not known.

RAND

Under PART, programmes are assessed and reassessed on a five-year schedule. The OMB acknowledges that in some cases this may be too short for results to be reflected in outcome/impact measures. Possible ways to deal with this problem suggested by the OMB are: to use output measures and/or measures ‘towards’ an outcome/impact.

Generally, outcome/impact goals are considered better than output goals. But using outcome/impact measures (to show that these goals have been achieved) is not unproblematic. Often it is hard to attribute changes in outcome/impact measures to actions of a programme (or research groupings in the HEFCE context). Two underlying issues that have been discussed in the literature are:

- Multiple causes of outcomes: Many of the outcomes sought by government programmes (research groupings) are subject to many separate causes
- Time lags: Some programmes (research groupings) have measurable outcome goals, but even when the programme (research grouping) is doing what it should, there may be a long time lag between programme (research grouping) activities and observable impacts.

An interesting way in which PART deals with these problems (which could serve as an example for HEFCE) is by leaving it to programme officials to come up with an appropriate solution and to make a compelling case why they think a programme has achieved its goal and/or annual target. Rather than prescribing a general way of how to handle attribution, this approach allows context-specific solutions and competition for good ideas.

Another way in which PART (indirectly) addresses the issue of attribution is by providing programmes with an incentive to choose outcomes/impacts that are not too difficult to attribute. To the extent that programmes are interested in influencing their performance assessment, making them decide upfront on their purpose, goals and corresponding measures, provides a strong incentive to choose outcomes/impacts that the programme can influence and which are likely to occur (at least in part) within a reasonable time frame. In other words, having to commit

to certain outcomes/impacts is likely to reduce the risk of programmes choosing outcome/impacts that are particularly hard to attribute.

PART is used across federal government. There is unlikely to be an issue with scalability. In terms of making sure that the evidence provided by programme officials is accurate, the OMB examiner in charge of a certain programme can audit a programme.

5. Case Study – Evaluating Research in Context

Context

- ERiC developed by a collaboration of higher education stakeholders in the Netherlands
- Developed criteria and indicators for assessment of societal quality / impact of research
- To be integrated into current system, Standard Evaluation Protocol (SEP) for second run 2009–2015
- Adapted from a system developed by network of science policy makers in the Netherlands, Sci_Quest, in the 1990s.

RAND

There is significant interest in developing approaches to assess broader impacts of research in the Netherlands. Traditionally, assessment has focused on academic rigour, but this was not felt to reflect the broad outputs of academic research across the portfolio (Spaapen, 2005). As a result, a number of approaches have been developed to assess impacts. These include the Leiden University Medical Centre societal impact evaluation and the use of the Payback framework by health research funder, ZonMW.

The current system for evaluation of HEIs in the Netherlands is the Standard Evaluation Protocol (SEP), which was established in 2003. This is a peer review system that evaluates research in four categories (VSNU, NWO and KNAW, 2003):

- quality (international recognition and innovative potential)
- productivity (scientific output)
- relevance (scientific and socio-economic impact)
- vitality and feasibility (flexibility, management and leadership).

Despite being explicitly stated as an assessment target, there is no methodology incorporated into this evaluation to assess impacts. This led to interest amongst the higher education community in developing a methodology for conducting impact assessment. As a result, the Evaluating Research in Context (ERiC) project was established in 2006 as a means to investigate impact assessment, with backing across the higher education field.

ERiC is adapted from a system developed by a network of science policymakers in the Netherlands, Sci_Quest, in the 1990s. A project group representing the full spectrum of higher education is responsible for organising ERiC's activities. The Netherlands Association of Universities of Applied Sciences (HBO-raad), the Royal Netherlands Academy of Arts and Sciences (KNAW), the Netherlands Organisation for Scientific Research (NWO), Quality Assurance Netherlands Universities (QANU) and the Rathenau Institute Association of Universities in the Netherlands (VSNU) are involved in the project, and Hogeschool Utrecht and the Ministry of Education, Culture and Science are observers. The mission of this project is to develop criteria, indicators and a methodology for assessment, as well as suggesting how this can be integrated into the next round of the SEP (2009–2015) (Spaapen, 2008).

Principles

- Provide a methodology for impact assessment which can be incorporated into SEP and is applicable across disciplines
- Comprehensive evaluation, focus on both quality and relevance
- Contextual: identify mission, involve stakeholders in indicator / benchmarking
- Combine quantitative and qualitative data
- Forward looking, learning rather than judging.

RAND

The main aim of the ERiC project is to provide a methodology that can be used for impact assessment. This should be flexible enough to be applicable across disciplines in a way that takes into account impacts in the arts and humanities that are often overlooked by traditional approaches. It is also desirable that this methodology be suitable for incorporation into the broader SEP assessment system.

ERiC is designed to be a comprehensive evaluation which focuses on both quality and relevance, and which is contextual, in that it takes into account the mission of the research grouping and incorporates stakeholder views in the benchmarking process. The methodology also uses a mix of qualitative and quantitative data to give the broadest picture of the outcomes of research. The end results of ERiC should be forward looking, focusing on learning rather than judging, and should be constructive in terms of the future path of the research institute (Spaapen, Dijstelbloem and Wamelink, 2007).

What it is: a multi-stage approach

Four-step process:

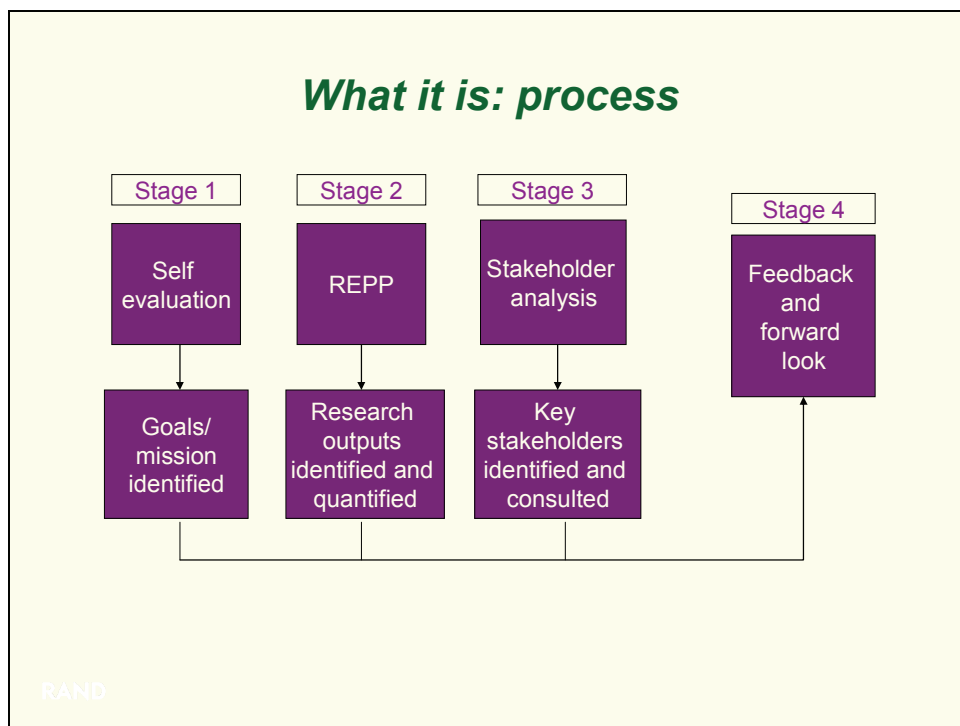
1. Establish self-image or mission of institute.
2. Data gathering to produce a research embedment performance profile (REPP). Domains used depend on research area.
3. Identify key stakeholders (using REPP and tracking trajectories of research outputs). Use questionnaire and/or telephone interviews to establish role of stakeholder and impact.
4. Feedback and forward look.

RAND

ERiC consists of four key stages (Spaapen, Dijkstra and Wamelink, 2007). The first of these is a self-assessment process in which the institute identifies its mission and/or goals. This can be achieved in a variety of ways, and different methods have been used in the pilot studies conducted. In one approach, the group constructed a self-image by scoring their orientation to the outside world on a 100 percent scale in different regions such as academia, industry, government, etc. Another approach followed a three-step process as follows:

- 1) The group estimates what percentage of their time they devote to work in different domains. This produces a task distribution on the basis of external orientation of the group and represents their self-image.
- 2) The group estimates the influence of stakeholders on the development of research in these domains. This establishes external influences on the research agenda and represents the contextual influence.
- 3) The most important stakeholders are established on the basis of a questionnaire and divided across the social domains. This establishes the stakeholder network or distribution.

This more complex methodology gives a more robust placing of the institute and feeds into the later stages. More recently, pilot studies have focused on rewriting the research group's existing mission statement in a SMART (Specific, Measurable, Agreed-upon, Realistic, Time-based) manner. Various other methods can be envisaged for establishing self-image and goals/mission.



The next stage is data gathering, with a focus on the research group’s performance in the various social domains, to produce a Research Embedment Performance Profile (REPP). The REPP is a map of the outputs of the group against various categories that may be externally developed for different disciplines, or may be developed in consultation with the researchers as part of the previous stage. These domains will necessarily differ depending on the research area. The number of categories can differ. In initial pilot studies, 30 indicators were investigated over five categories: science and certified knowledge; education and training; innovation and professionals; public policy; and collaboration and visibility. This was simplified to 15 indicators in three categories (science/certified knowledge, industry/market, and policy/societal) in a subsequent pilot study to save time whilst still preserving sufficient information to give an insightful assessment. Data gathering relies partly on good record keeping by the research organisation.

The third step identifies productive interactions with industry, policy and society at large and maps external stakeholders. These may include collaborators, end users and non-funding organisations with a general societal mission related to the research conducted. The stakeholders are identified using the REPP and by tracking trajectories of research outputs. A number of stakeholders are then contacted, either by survey or telephone interview, in order to establish their role in research (eg do they have any input into research direction or dissemination methods) and the downstream impact of the research.

The process has developed during the course of pilot studies, with both survey and interview methods found to offer distinct advantages. A survey is advantageous in that it obtains opinions from a wide range of stakeholders that would not be possible through telephone interviews due to time constraints. However, a lot of additional insight is found to come from interviews, as views are not constrained to the survey template, allowing interactions, which are often complex, to be better understood. The best approach is thought to be a mix of both survey and interviews to obtain broader information as well as some deeper insight (Spaapen, Dijkstra and Wamelink, 2007).

The final stage consists of feedback and a forward look. The outputs from stages two and three are compared against the mission/goals established in the initial stage in order to illustrate the level of match or mismatch between the group's self-perception and its true impact. On this basis, future plans are made that can build on established strengths and target identified weaknesses.

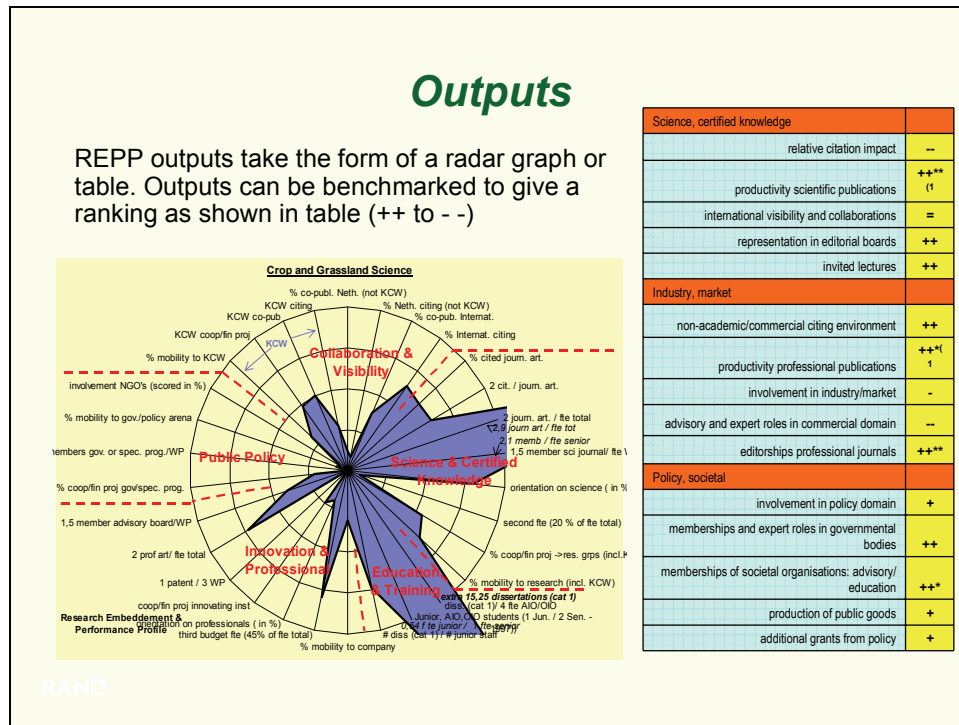
Where is it being used?

- Who developed it? Originally developed by Sci_Quest, now being developed by NWO and collaboration across Dutch higher education system
- Who is using it?
 - Pilot studies including TU Delft Architecture and VU Amsterdam Law
 - To be incorporated into SEP 2009–2015 period.
- Also being developed as 'SIAMPI' in collaboration with European partners with case studies in health.

RAND

ERiC was originally developed by Sci_Quest, but is now being developed by the Netherlands Organisation for Scientific Research (NWO) with support across the Dutch higher education system as outlined previously. Pilot studies have been conducted in Agricultural Sciences at Wageningen University and in Pharmaceutical Sciences at the universities of Utrecht and Groningen. These studies produced the REPP profiles shown in the next slide. Further case studies are ongoing at TU Delft Architecture and VU Amsterdam Law, which are intended to establish the method for use outside scientific fields, for which it was originally developed (*ERiC project: www.eric-project.nl*). There are plans to incorporate parts of the ERiC methodology into the SEP for the 2009–2015 period (Spaapen, 2008).

The approach is also being developed as 'SIAMPI' (a project studying Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions between science and society) in collaboration with other European partners. Here, case studies are being conducted in the Health field at UMC Leiden University, UMC Free University Amsterdam and NIVEL, a primary health care organisation (Spaapen, 2008).



The REPP can take the form of a radar chart or a table. These express the level of activity of a group in different fields on a scale. In the case of the radar chart, this is against a scale increasing radially from the centre of the chart. For the table, this is a five point scale from -- to ++. The baseline against which impacts are scaled varies, but is typically the ‘average’ level of impact for a research institute in this field. This can be difficult to establish in the first instance but may emerge more clearly over time as the methodology is used repeatedly. These ratings can also be used as a form of benchmarking, as they compare research outputs against a standard or desired level.

In the initial pilot studies, a radar chart format was used. In this case, five areas were analysed, but this presentation can be used for fewer categories. This format gives an instant pictorial impression of the outcome of research across areas. However, it was felt that the order in which categories were placed around the graph could have an influence on the way in which data was perceived. More direct comparison between adjacent segments is inevitable, and the appearance of the graph can differ substantially depending on the way in which it is put together. This was considered undesirable, though it would be possible to use this method of presentation in a slightly different format, such as a histogram (Spaapen, Dijkstra and Wamelink, 2007).

In a later pilot study a table format was used. Though the table format is perhaps less visually appealing than a radar chart, the interpretation is less dependent on the order in which categories are placed. A table is better suited to a reduced number of categories, such that the volume of material is manageable in that format.

How it is being used

- Consequences:
 - There is no direct link to funding
 - Dissemination channels not yet defined but likely to include report on institute website.

RAND

In the Netherlands, the outcomes of this type of assessment are unlikely to have a direct impact on funding decisions in the short term. Rather, findings are intended to guide improvement of the research programme and future development. However, benchmarking is being incorporated into the approach such that it would be possible to convert the results into a ranking that could be used for funding decisions.

It is not yet clear how the results of this assessment will be disseminated. However, it is most likely that results will be disseminated as a final report available through the HEI, and possibly also the assessment organisation, probably via their website(s).

Stakeholder involvement

- Stakeholders:
 - System is run by evaluator who conducts survey/interview and gathers REPP data
 - Mission/self-image defined by institute
 - Users respond to interview/survey
 - Some reliance on good record keeping by institute for REPP
 - Not clear how interim self-assessment in SEP will be organised
 - Positive feedback at pilot stage.

RAND

Evaluatees play a significant role in the ERiC assessment process. This falls largely within the first stage, in which the self-image of the research institute is established and a mission/goals set. An evaluatee questionnaire can also form part of the data gathering for the REPP, and thorough record keeping at the research institute can significantly reduce the burden of data collection in stage two.

Evaluators are involved in all stages of the assessment. In the first stage, they work with the research institute to help establish a mission and/or goals. In the next stage, they are responsible for selecting appropriate REPP categories, which may be a significant challenge in the first instance but is likely to be less difficult in subsequent assessments. Data gathering is then relatively straightforward provided appropriate records have been kept. In stage three, evaluators identify and map key stakeholders. They will also be responsible for the composition and analysis of the survey and/or conducting telephone interviews. Again, the burden here should decrease with subsequent evaluations as a range of stakeholders are already identified and protocols for surveys and/or interviews are established. In the final stage they are involved in feeding back their findings to the institute.

Other stakeholders (eg collaborators, users, etc) are also involved in the process through their participation in a survey and/or interviews in stage three.

It is not yet clear if the process will be conducted at the three-yearly self-assessments in the SEP or only in conjunction with six-yearly external assessments. If the process is to take place every three years, it is unclear how it will take place.

From the pilot studies, there has been a good level of stakeholder buy-in. However, ERiC has not been widely tested and it will be interesting to gauge opinion if elements are included into the SEP later this year. The methodology has support across higher education stakeholders as shown by the wide range of backers for its development. There is particularly strong support from the arts and social sciences who believe an assessment of this type will meet their needs better than traditional approaches.

Additional technical aspects

- **Currently at Pilot study stage but to be incorporated into SEP as methodology for impact assessment strand.**
- **Attribution: not addressed directly. Input of users may help to address this.**
- **Time frame: self assessment every three years, external review every six years. Flexible.**
- **How many submissions: assesses quantitatively across all impacts. Qualitative assessment coverage depends if survey or interview format.**
- **Additionality: not addressed. Not easily included.**
- **Scalability: not addressed. Potentially challenging.**
- **Application across areas: initially developed for sciences but designed to be broadly applicable. Pilot studies in range of areas.**
- **Aggregation: institute - 'organisational unit covering a more or less coherent area of research'.**
- **Decision making: not currently used for funding decisions. Benchmarking in REPP being investigated.**
- **Auditing: stakeholder interviews. Likely to be part of a larger peer review in SEP in first instance.**
- **Burden: not yet clear. Likely to be heavy in first instance as REPP categories established, decreasing in subsequent assessments.**

The slide above outlines a variety of technical aspects with regard to the operation and implementation of the approach. In black, we outline how the approach is currently being used, with the pink text summarising the issues in extending the approach to a nation wide research impact evaluation approach. We discuss each issue in more detail below.

As outlined previously, ERiC is currently at the pilot study stage, but may soon be incorporated into the SEP. If so, this is conducted as self assessment every three years, with a further external review based on this every six years (VSNU, NWO and KNAW, 2003). How ERiC would be used at the self-assessment stage is not clear, but in terms of timescale, the approach is flexible.

ERiC is designed to be applicable across all fields, although it was initially developed for application in the sciences, and pilot studies have been conducted across a range of fields. The REPP categories and group mission are constructed specifically for each discipline. However, this has associated problems as it is likely to result in a high level of burden, particularly for the first round of assessments. The burden is likely to decrease with subsequent iterations. The method can also be used to assess all impacts, though different approaches have been used, such as limiting the number of academic papers submitted to a 'representative' sample of 15, in order to reduce the burden (Spaapen, Dijstelbloem and Wamelink, 2007). The level of coverage will also depend on whether a survey or interview approach is used to assess stakeholder involvement.

Assessment is at the level of a research institute, which is defined as 'an organisational unit covering a more or less coherent area of research' (Spaapen, Dijstelbloem and Wamelink, 2007). This allows for the selection of REPP categories that are appropriate for all impacts created. Assessment at the HEI level, where a range of different disciplines are included, would be difficult, except through aggregation of scores gained on the departmental or 'institute' level.

In the Netherlands, evaluation is not currently tied to funding allocation. However, benchmarking can be incorporated into the REPP, which would allow ranking to take place. Again, in the initial stages, this will take a significant amount of effort, and may be less robust until 'typical' levels for comparison can be established over several assessments.

The role of stakeholder interviews may be significant in terms of attribution of impacts, as well as verification. Engagement with stakeholders, particularly end users, is a further issue to be considered. If ERiC forms part of the SEP, the outputs are likely to be subject to some element of peer review, which will also ensure submissions are verified.

Additionality is not considered in ERiC, although the goals/mission set by each institute could be selected to address this to some extent. Scalability is also an issue for this approach which has not been clearly addressed. The higher education system in the Netherlands is significantly smaller than in the UK so the issue may be less pertinent for application in the SEP than the REF.

6. Issues and Ideas – Applying Lessons for the REF

Issues and analysis

Drawing on the case studies, key issues were identified and analysed as follows:

- traffic lighting
- identifying issues individually
- clustering issues
- exploring clusters in more depth.

RAND

At a workshop of the RAND Europe team and the HEFCE project team, the four frameworks were assessed against HEFCE's criteria of an adequate impact framework on the basis of the case studies. For this purpose a *traffic lighting* scheme was used – with a green light if a framework meets a certain criterion; amber if the criterion is met in part; and a red light if a framework does not meet the criterion in question.

In addition, using the learning from the four case studies, issues were identified and analysed. This involved two main steps:

- **Identifying issues:** Issues were identified independently by individuals.
- **Clustering issues:** Issues were clustered into areas in discussion with HEFCE at the workshop.

Through clustering, a number of key areas were identified. These were then investigated in greater depth. In this section we outline the learning from this process, identifying a number of key observations that may inform future work in this area.

Mapping the case studies against HEFCE criteria

	Credible & acceptable to academic and user communities	Encompasses full range of impacts (economic, social, public policy, welfare, culture and quality of life)	Can be adapted to apply to all disciplines (within a single broad approach)	Is practical and does not generate excessive workload	Avoids undesirable perceptions and incentives
RQF (Australia)	G	G	G	A	A
ARC (UK)	A	R	R	A	A
PART (US)	A	A	G	A	A
ERiC (Netherlands)	A	G	A	R	A

RAND

The above chart ‘traffic lights’ the four case study frameworks against five criteria identified by HEFCE. The green, amber or red assessment was determined at the workshop, by the HEFCE project team and the RAND Europe team. It thus reflects the collective subjective view of the workshop participants.

It is evident from this analysis that the Australian RQF provides the ‘best fit’ against the emergent criteria for REF. This also corresponds with our judgement in undertaking this work. Hence our first key observation is that *the work of the Australian RQF Working Group on Impact Assessment provides a promising basis for developing an impact approach for the REF.*

There are a number of other observations to be made from the above analysis. First, none of the frameworks rate well on the last two criteria – that is the burden of assessment and lack of undesirable incentives. Second, there is some learning from the PART and ERiC frameworks. The former provides an element of self-assessment against previously self-set mission and objectives. This could be applied to the REF, especially in the context of the ‘environment’ module (which is not within the scope of the current report). The ERiC framework provides a comprehensive, but high burden, measurement of impacts and some of these could provide illustrative examples of verifiable evidence for case studies. Finally, there is evidence from within the RQF and RAISS that *impact indicators are not yet sufficiently developed and tested to be used to make funding decisions.*

Identifying issues from the four case studies

On the basis of an issues analysis six key issues with the four case studies could be identified:

- Strategic intent
- Definitions (ie a common language)
- Unit of analysis
- Attribution
- End users
- Burden

Each of these six areas is discussed in further on the following slides.

To draw out potential lessons for developing an impact approach for the REF, at the workshop, the RAND Europe project team and the HEFCE project team identified issues and ideas coming through from the case studies.

To identify the issues, individuals were asked to identify the five most important issues arising from reviewing the case study frameworks. The issues were then discussed and, on the basis of this discussion, clustered. Six core themes arose:

- **Strategic intent.** This included: ‘linearity of funding’, ‘perverse incentives’, ‘efficiency (inputs versus outputs)’ and ‘path-dependency’ (that is, risk of a positive assessment leading to a higher chance of a positive assessment in follow up evaluations, and so on)
- **Definitions (ie a common language).** This included: ‘identifying what impact means for different disciplines’, ‘definitions/expectations of impacts by discipline’, ‘definitions of impact’, ‘differences between disciplines’, ‘applicability across disciplines’ and ‘scalability’ (of case study framework)
- **Unit of analysis.** This included: ‘granularity, groupings, departments, HEIs’, ‘unit of assessment’, ‘aggregation – unit of assessment used’
- **Attribution.** This included: ‘researchers who move’, ‘shared impacts?’, ‘impacts can ‘grow’ in time’, ‘time frames for impacts’
- **End users:** This included: ‘involvement of users in process’, ‘identifying and engaging wider stakeholders’, ‘involvement of end users’, ‘will end users engage?’
- **Burden.** This included: ‘who has to do the work?’, ‘verification’, ‘can REF impact information be used for other purposes’, ‘information overload for panels’.

Each of these six areas is discussed in further depth on the following slides.

Being clear on strategic intent

- Some of the intuitive appeal of linking budget decisions to performance is that it gives an opportunity to improve alignment of the objectives of research funders and researchers/research groups/research institutions.
- In order to get incentives right, however, a number of aspects need to be considered:
 - HEFCE needs to decide what functional form should underlie its funding regime.
 - It also needs to take into account the ‘noisiness’ of its performance measures and the risk aversion of researchers/research groups/institutions.
- Linking budget decisions to performance also raises some dynamic issues which HEFCE needs to consider. The most important ones are:
 - HEFCE needs to be explicit about its take on path-dependency and concentration effects.
 - HEFCE should take into account the special situation of young researchers/research groups/institutions.

RAND

One of the challenges of any allocation system is to link performance to funding in the intended way in a transparent manner. This was a challenge for the RAE and will be one for the REF. Arguably, this challenge becomes greater when assessing research impact, given the diverse nature of impacts and the time lags involved. *To be effective it will be essential that the approach can provide an unambiguous rating and comparison of submissions.*

It will also be important to determine how the assessment of impacts relates to the allocation of funding. This could either be a relationship where all assessed impacts are rewarded or a relationship that aims to concentrate funding in areas of excellence (as with the current RAE).

An additional factor in determining the overall funding an institution receives will be the relative weight given to the three assessment profiles: quality, impact and environment. In the ERiC system institutions are allowed some flexibility in how they weight on different indicators. This raises the possibility that *HEFCE could consider allowing some flexibility in the weights given to the three areas of assessment – quality, impact and environment.* For example, HEFCE could announce that impact would carry between 15 and 25 percent weighting, quality between 50 and 70 percent and environment between 15 and 25 percent. This would allow, and incentivise, HEIs to articulate and focus on a clear research vision and reward success against that vision; and would also recognise the diversity in the sector.

Finally, *because the REF will drive the behaviour of HEIs, it will be important to be very explicit about the criteria for assessing impact.*

Have a clear understanding of what is meant by 'impact'

- Central to the attempt to evaluate impact is to define what is meant by impact.
- The frameworks reviewed provide some examples of possible definitions. To the extent, however, that the definition of impacts will have significant behavioural implications, HEFCE may want to come up with its own definition (tailored to the present context).
- For its own definition HEFCE needs to take into account a number of aspects – including:
 - Who should define what is 'impact' – researchers/research groups/institutions or HEFCE itself?
 - When should impacts be defined – before or after the evaluation period?
- The main issues with regard to the measurement of impacts are whether objective or subjective measures should be used and whether these measures should be relative or absolute measures.

RAND

The definition and measurement of impacts raise a number of issues. The first one is who defines what is meant by impacts? In the PART framework programme officials define the purpose of their programme, its long-term (impact) goals and annual targets. In the case of the RQF and RAISS, on the other hand, the research funder/evaluator defines what is meant by impact. ERiC is a hybrid in so far as the definition of impact (in terms of impact categories) is developed in collaboration of researchers and research funders/evaluators.

The advantage of involving researchers in the definition process is that it allows for more diversity – and decreases the probability that an important dimension is excluded from the definition. On the other hand, leaving the definition of impacts to researchers/research groups/HEIs bears a number of risks:

- It may lead to the inclusion of impacts HEFCE does not consider particularly valuable (and consequently may steer research into areas HEFCE is not concerned with)
- It may result in a host of different definitions of impacts, and therefore comparison of different researchers/research groups/HEIs may be more difficult
- If impacts are defined after the event, there is a danger that researchers will most highly value the impacts they have actually produced, rather than other impacts that they might have produced and which would be more valuable to society.

To balance these issues HEFCE should consult widely and then provide a definition of how it defines impact.

The measurement of impacts raises a number of further questions. The first one is whether to use quantitative or qualitative measures. The idea of quantitative performance evaluation is to evaluate performance against a set of metrics. Its biggest drawback is that it is likely to encourage researchers/research groups/HEIs to game the evaluation process to their advantage.

Metrics are typically imperfect. They tend not to include all activities that are desirable; and nor can they differentiate between when a particular activity is desirable and when it is not. As an

example, trying to get research into policymaking – even if typically desirable – can be undesirable if, for example, the research is not robust or has severe ethical implications.

However, qualitative assessments – such as case studies – are likely to also induce inefficient behavioural responses. The literature in both economics and human resources management has emphasized how incentives provided through qualitative assessments can cause those being evaluated to change their behaviour and evaluators to distort their reports (eg grade inflation) in such a way that it harms efficiency.

Because of the imperfections of both quantitative and qualitative measures, HEFCE should use a combination of the two, such as using case studies or narratives supported by possible proxy indicators of impact as proposed for the RQF.

A final measurement issue is whether there should be an absolute evaluation of impacts or a relative one. Should HEFCE evaluate impact against one scale across all fields – meaning that the best performing grouping in one subject areas would score much lower than the best performing grouping in a unit of analysis that generates more impact; or should HEFCE evaluate the performance within different fields – so that the best grouping in each unit of analysis gets the highest possible ranking, no matter how it fares relative to groupings in other subject areas.

Have a clear idea of the unit of analysis

- Evaluation can take place at different levels of aggregation – ranging from low (individual researcher) to intermediate (research group, faculty) to high (university or discipline).
- The best level of aggregation for assessing impacts is likely to be influenced by two countervailing forces:
 - Principal-Agent theory favours lower levels of aggregation – to reduce the free-rider problem.
 - Evaluation practice favours higher levels of aggregation.
- In addition to the level of aggregation HEFCE should also take organisational aspects into account when defining the optimal unit of analysis.
- HEFCE should try to avoid to break up existing formal and informal organisational structures.

RAND

Having looked at the question ‘What to measure?’, we can now turn to the next question ‘At what level of aggregation to evaluate?’. The level of aggregation in the REF could be low (individual researcher), intermediate (research group or faculty) or high (HEI or discipline).

The RAE used subject-based Units of Assessment. The complexity for assessing impact is that impact groupings may not align with departmental or institutional groupings. And researchers in the same subject area may have different types of impact. However if impact assessment is going to include a qualitative case study based strand the burden of assessing at a very low level of aggregation is likely to be too high. Furthermore, evaluation studies tend to favour higher levels of aggregation if impacts are to be assessed (Brutscher et al., 2008). The reason is that, because impacts tend to occur after a long time lag only and are typically the result of a number of contributors, it is more difficult to attribute a change in impacts at low levels of aggregation than high levels of aggregation.

An attractive way to address these issues is the *RQF approach which allows researchers and institutions discretion in how they organise into groupings – while still requiring groups to have some minimum size*. It would also allow researchers to form groupings that crossed existing formal or informal organisational structures, but still remain within a cognate discipline.

Acknowledge the challenge of attribution

- Attribution is the ascription of a causal link between observed changes and a specific intervention. In the context of research assessment, this means demonstrating a link between original research and verifiable impact.
- It is widely acknowledged in the research evaluation literature as one of the biggest challenges for research evaluation and assessment. As the recent Canadian Academy of Health Science report points out, there is both a risk of over-attribution and under-attribution:
 - Over-attribution would occur when multiple research groups claim the same impact, thus risking double counting
 - Under-attribution would occur when the researcher is unaware of the impact that they have had
- There are two practical ‘work-arounds’ to these problems:
 - Examine ‘contribution’ as opposed to ‘attribution’.
 - Be explicit about using a sampling approach where the type of benefits are illustrated, as opposed to the totality of impact being measured.

RAND

One of the biggest issues in relation to the assessment of research is that of attribution. This is an issue that has been raised in a number of research evaluations (eg Wooding et al., 2005) and was examined in some depth in the recent Canadian Academy of Health Science (CAHS) report, *Making an Impact* (CAHS, 2009). The only two ‘solutions’ offered by the CAHS are to: (a) undertake multivariate regression analysis to partition impact between its various causes based on various explanatory variables; this approach has been used at the macro level to examine the influence of research on economic growth in 21 OECD countries between 1971 and 1998; or (b) to build up case studies that examine the research antecedents to a set of impacts; this approach has been used in a number of historical studies (eg Cromroe and Dripps, 1976) but is too expensive and time consuming to be practical for a REF.

For the REF there are two key concerns – firstly the necessity for an accurate assessment of attribution to allow a fair allocation of credit for a particular impact and secondly the need to verify such assessments of impact.

To address the first issue, the approach proposed for the RQF provides a possible way forward. This is to *acknowledge that the issue of attribution is complex, but can be mitigated by a case study approach, and by focusing on those impacts where attribution is more readily demonstrated. It needs to be made explicit that case studies are illustrative, rather than intended to measure the total impact of an HEI.*

Secondly, *a verification mechanism is likely to be important in building confidence in the approach* and could be addressed by an auditing framework that would review impact claims and could double check a sample of impacts.

Undertake further work to understand the degree of administrative burden for different stakeholders

- One of the overarching aims of the proposed REF is to “to reduce the administrative burden on institutions in comparison to the RAE”
- As illustrated in a preceding slide, none of the frameworks we looked at rated well in terms of not generating excessive workload
- Indeed given the observation that review based mechanisms will be needed to assess impact, it is clear that there will be an administrative burden
- This burden will fall to two different stakeholders groups:
 - The HEI preparing impact statements
 - The panels assessing impact statements
- HEFCE will need to undertake further work to model what the likely burden of assessing impact will be, and compare it against previous RAE

RAND

One of the reasons that the REF was proposed was to reduce the transaction costs of the RAE. Thus, it was anticipated that the REF would ‘produce robust UK-wide indicators of research excellence for all disciplines’ and that by using an indicator-based approach the administrative burden on HEIs would decline.

It is clear from the four case study frameworks that have been reviewed that the administrative burden of assessing impact is relatively high – three of the four frameworks were rated amber, and the other (ERiC) red in the assessment in section 5.

If expert-based review is to be used for impact assessment in REF (as seem likely from this work), then the burden will be two-fold: HEIs will need to prepare submissions demonstrating their impact, and these submissions will need to be reviewed by panels.

The burden related to preparing impact submissions is likely to have a high initial cost as this will be the first time the HEIs have undertaken such an activity; however, there are some indications that for the RQF the universities that had moved to the pilot stage of preparing submissions have subsequently found this information helpful for many other purposes such as fund raising and lobbying. In this context there is considerable learning that could be gained from unofficial Australia pilot work, notably the work of the Innovative Research Universities Australia. As discussed overleaf, the consequence of the burden for panels is more of an unknown as it will be necessary to engage non-academic reviewers in assessing impact.

In other words, the issue of administrative burden – and how the burden of the REF will compare to previous RAEs – is a topic that needs further work and we suggest that *HEFCE will need to undertake further work to understand the likely burden of assessing impact, and compare it against the burden of previous RAEs.*

Clarify the role of end users, and what is expected of them and whether this is realistic

- A recurring issue in developing any qualitative assessment of impact is the need to engage end users of research. This is the case for the RQF and ERiC.
- The role of end users is significantly more important in assessing research impact than research quality, as it is the sole way of verifying and auditing claims of impact made by researchers.
- However, as end users have no 'stake' in the outcome of the assessment process, it is potentially hard to recruit individuals especially given the likely time commitment.
- This is not a new issue. In RAE2008 additional panel members with experience of using research could be recruited for the main panels:
 - It is recommended that HEFCE reviews how successful this was, for understanding motivations of non-academics for participating or not.
- The RQF was welcomed by the business community in Australia, as it was felt it would lead to more 'relevant' research; however commitment was never tested as the system was not rolled out.

The engagement of end users of research will be critical to the success or otherwise of the REF. Research quality is verifiable and auditable through either peer review or bibliometrics. There will be a similar need to verify claims of impact in submissions by HEIs, especially if a system is based on the RQF. As noted earlier there are no robust indicators for research impact similar to bibliometric analysis of research publications and citations. Thus the only alternative is some form of review. As the focus will be on research impact – as opposed to research quality – it will be important to recruit reviewers who can make an independent assessment of high impact research versus, for example, low impact research. This will mean identifying reviewers who are the end users of research, such as business people, policymakers and others.

Such end users, however, have no reason to participate in such a review. The *quid pro quo* for HEIs is that they let their academics engage freely in previous RAE in the knowledge that the process leads to funding being directed to the HEI. Beyond contributing to the public good, there are no such 'deals' for non-academic based reviewers.

HEFCE needs to do further work to understand the motivations, likelihood and commitment of end users in participating in the assessment of research impact. This could include speaking to organisations such as the CBI, analysis of the motivations for non-academic panel members and assessors from previous RAEs (including those who turned down the opportunity) and estimating the likely commitment for such reviewers.

Appendix A: Long List of Frameworks

The following table provides an overview of frameworks we considered in the first part of the project. Please note that the ‘issues’ column summarizes what was stated in the literature – not our own analysis. In addition, to the extent that our review of the literature around the frameworks was not systematic, the issues should therefore be considered indicative rather than conclusive.

Approach	Where used	Assesses wider impacts?	Main methodologies	Description	Issues
RAE	UK, withdrawn after ~20yrs use	No	Portfolio, peer review, benchmarking	Portfolio submitted, assessed by expert panel for output and research environment/esteem indicators. Rating on 1-5 scale published.	<ul style="list-style-type: none"> - Expensive and unpopular - Compromised clinical academic research as teaching/clinical practice not valued - Perceived improvements may be from academics’ “games playing” - 55% now in top 2 grades - little discrimination - Problem exacerbated by UK Government failure to fully fund improved ratings in 2001 cycle.
RQF	Proposed in Australia, but never used	Yes	Portfolio, peer review, benchmarking	Research quality measured on a 5 point scale, impact on a 3 point scale to give overall rating. Conducted through portfolio generation then peer review as for RAE	<ul style="list-style-type: none"> - Diminished independence of research councils if RQF governs funding allocation, some groups may be ineligible - Expensive
PBRF	New Zealand, since 2003	No	Portfolio, peer review, benchmarking	3 elements: <ul style="list-style-type: none"> - to reward and encourage the quality of researchers: 60 percent of the fund - to reflect research degree completions: 25 percent of the fund 	<ul style="list-style-type: none"> - Devaluation of teaching - Reduced academic autonomy - Disadvantages applied research - Discourages

				<p>- to reflect external research income: 15 percent of the fund</p> <p>Portfolio assessed by expert panel for quality of their research outputs, the esteem they are held in by their peers and their contribution to the research environment.</p> <p>Score 0-7 for each category.</p>	<p>collaboration</p> <ul style="list-style-type: none"> - Cost-benefit ratio – many institutions find exercise costs more than resulting gains in funding - Staff individually categorized – many early career researchers very low rankings – negative impacts on staff morale
ERA	Proposed in Australia in place of RQF	Yes	Portfolio, bibliometrics, benchmarking(?)	<p>Assess research under 3 categories</p> <ul style="list-style-type: none"> - Research Activity and Intensity e.g. research income, number of PhD students and completion rates, staff levels - Research Quality eg Publication analysis and research grants awarded by peer review - Applied Research and Translation of Research Outcomes: This will differ between disciplines – advice to be sought. May include patents, commercial revenue, performances/exhibitions, presentations, etc. 	
PART	US, around 800 federal programmes	Yes	Questionnaire, benchmarking, peer review	<p>Diagnostic questionnaire, with four critical areas of assessment – purpose and design, strategic planning, management, and results and accountability.</p> <p>The first set of questions gauges whether the programs’ design and purpose are clear and defensible. The second section involves strategic planning, and weighs whether the agency sets valid annual and long-term goals for programs. The third section rates agency management of programs, including financial oversight and program improvement efforts. The fourth set of</p>	<ul style="list-style-type: none"> - Consistency – answers to different question may depend on interpretation - Defining adequate performance measures - Minimising subjectivity - Measuring progress towards results e.g. in early stages of a programme - How to do this regularly without excess cost - Assessing overall context –

				<p>questions focuses on results that programs can report with accuracy and consistency.</p> <p>The answers to questions in each of the four sections result in a numeric score for each section from 0 to 100 (100 being the best). These scores are then combined to achieve an overall qualitative rating that ranges from Effective, to Moderately Effective, to Adequate, to Ineffective. Programs that do not have acceptable performance measures or have not yet collected performance data generally receive a rating of Results Not Demonstrated.</p>	<p>programme does not account for complementarily between programmes</p> <p>- How adaptable to a university assessment system?</p>
MORIA	Not yet used, developed in Australia	Yes	Questionnaire, benchmarking, bibliometrics	<p>Measures outputs of health research across three domains:</p> <ul style="list-style-type: none"> - Knowledge - Health gain - Economic benefits <p>Developed as analytic instrument in peer review process for grant applications, building on Record of Research Achievement framework. Findings translated into numerical score for activity, recognition and impact in all 3 categories.</p>	
Payback	CIHR, Canada (but also UK DH, ESRC, and others on ad-hoc basis)	Yes	Questionnaire, case studies, bibliometrics, logic modelling	<p>Framework to measure outcomes of health research consists of a definition of evaluation criteria and a logic model. Conducted through case studies.</p>	
Societal Impacts of Health Research	LUMC, Netherlands (hope to extend across Netherlands once)	Yes	Questionnaire, benchmarking.	<p>Societal impacts of health research (to address previous imbalance of Netherlands towards scientific quality rather than impact on society). Findings translated into a</p>	- Societal impact only

	electronic data collection system ready)			numerical score. Categorise as knowledge products, exchange/esteem, use and attractiveness, each of which measured across private sector, public sector and general public.	
Various	Vinnova, Sweden	Yes	Interviews, case studies, macroeconomic modelling, microeconomic modelling, productivity analysis, peer review, bibliometrics	Impact analysis of Vinnova's work conducted annually. Consists of ongoing evaluation process and an impact analysis. Have tried several methods for impact analysis, including interviews, case studies, and socio-economic analysis. Useful in talking about advantages and disadvantages of different methodologies.	
Economic Impacts of Investment in research and Innovation framework	DIUS, UK	Yes	Questionnaire, benchmarking, macroeconomic modelling, productivity analysis, interviews and case studies	Models delivery of economic impacts at the aggregate economy level through 3 stages – innovation outcomes and outputs, knowledge generation, and investment in the research base and innovation	- Economic impact only
Framework Programme	EU	Yes	Questionnaire, macroeconomic modelling, interview and case studies, cost-benefit analysis, peer review.	Track results of research and match to policy goals to identify what needs to be improved to improve this match. Interim and ex post evaluation by expert panels. Evaluation methods include sampled analyses, case studies and longitudinal surveys and where appropriate cost-benefit and/or macro-economic impact analysis.	
CDMRP	US army medical research and material command	Yes	Interviews and case studies	Manages biomedical research. Consists of several elements, incl. grant management system, product database and Concept Award Survey. Product database tracks products of research and stage of development against a number of criteria to map development and track where promising	

				impacts/products are stalled or stopped.	
SEP	Netherlands	Yes (in theory)	Self assessment, peer review	<p>Assesses research on four main aspects, namely:</p> <ul style="list-style-type: none"> • Quality (international recognition and innovative potential) • Productivity (scientific output) • Relevance (scientific and socio-economic impact) • Vitality and feasibility (prospects for the future) <p>Self assessment report followed by site visit and openly published report. Ranked, but does not directly decide funding. Conducted as a two-way dialogue between evaluator and evaluatee. Supposed to include element of societal impact assessment but criticised as not doing it well. Programme to be re-assessed 2009 and may adopt elements from ERiC which is well regarded in the Netherlands.</p>	
ERiC	<p>Netherlands – pilot studies</p> <p>TU Delft Architecture and VU Amsterdam law.</p> <p>Developed by Sci_Quest.</p>	Yes	Bibliometrics, survey, interviews	<ol style="list-style-type: none"> 1. Rewriting the research group's mission in an assessable (SMART) manner. SMART stands for Specific, Measurable, Agreed-upon, Realistic, Time-based. 2. Collecting data for the purpose of the indicators. The data are qualitative or quantitative, for example attractive descriptions of research results. A profile of the group can be compiled based on these data. 3. Relevant stakeholders are interviewed, or are sent a questionnaire by e-mail, about the relevance 	

				<p>and quality of the research. During this stakeholder research every effort is made to ensure that the results reflect the diversity of the relevant parties. (This could also be via interview)</p> <p>4. In this final stage, the results from the first three stages are drawn together in a report. This report can be used to reflect upon the extent to which the intended quality and relevance have been realised</p>	
CNRS	France, under review	No	Questionnaire/por tfolio(?), peer review	<p>Researchers write activity reports every 2 years with list of publications.</p> <p>Range of activities included: scientific references, mobility, opening up to industry, teaching activity and dissemination of the scientific culture. Sections also provide opinion on requests for promotion and reallocation requests, etc. Every 4 years, section re-assesses whether re- search units and candi- dates in line with frame- work and strategy of CNRS. For analysis on activities of unit, section relies on quadrennial report of committee of the French National Agency for the Evaluation of Research and Higher Education (AERES) evaluation.</p> <p>Process undergoing reform</p>	- Internal rather than comparative – not useful for funds allocation
PNECU	RES, Spain	Possibly, though not a focus	Questionnaire, interviews, peer review	<p>Mixed system of self- assessment and external assessment (through a visit with interviews), followed by writing and publication of a final report. Although more oriented to teaching activities than to research ones, research activities in universities partially</p>	- Internal rather than comparative – not useful for funds allocation

				reviewed, paying attention to different institutional dimensions: scientific production, external relationships, human resources, support staff, economic resources, material resources and infrastructures, doctoral programs, research groups, university's internal support for research, promotion activities, internal communication and collaboration, and departments' internal promotion of research.	
Higher Education Impact Model	Library House, UK	Yes	Questionnaire, benchmarking	Used to assess impacts of Cambridge University on an economic and societal level. Developed to be applicable across all Higher Education Institutes. Provides list of criteria to look at in similar way to payback framework.	
NIAD-UE	Japan	Yes	Questionnaire, bibliometrics, peer review	Each institution submits a performance report; evaluation is based on document analysis and site visits. Evaluation based on objectives of university, criteria are organisational structure, research policy and programme, research quality, social/economic/cultural impacts, internal quality improvement system.	- Internal rather than comparative – not useful for funds allocation
CONEAU	Argentina	Depends on internal goals of institution	Questionnaire/portfolio, peer review	Peer review process including site visit based on self assessment completed internally. Does not produce rankings etc. Only used for internal improvement. Recommendations against universities own internal targets.	- Does not rank – will not be useful in funds allocation

Appendix B: References

- Brutscher, P-B., S. Wooding and J. Grant. 2008. Health Research Evaluation Frameworks: An International Comparison. RAND Europe.
- Cromroe, J. H. and R. D. Dripps. 1976. Scientific basis for the support of biomedical science. *Science* 192 (4235):105–111.
- Donovan, C. 2005. A review of current Australian and international practice in measuring the quality and impact of publicly funded research in the humanities, arts and social sciences *REPP Discussion Paper 05/3*.
- ERiC project website*. Available from <http://www.eric-project.nl/>.
- Expect More website*. Available from <http://www.expectmore.gov>.
- Gilmour, J. B. 2006. Implementing OMB's Program Assessment Rating Tool (PART): Meeting the Challenges of Integrating Budget and Performance.
- Kniker, T. 2006. Comment to W. Trochim on An Introduction to PART (2006). Available at http://www.eval.org/part_3.asp as of 22/04/09
- Medical Research Council. *Outputs Data Gathering Tool* 2008 [as of Tues 21 April 2009] Available from <http://www.mrc.ac.uk/Achievementsimpact/Outputsoutcomes/index.htm>.
- OMB. 2005. Guidance for Completing the Program Assessment Rating Tool (PART).
- Peacock, J., L. Harvey, M. Barber, P. McFadden, W. King, T. Enright, D. Schreuder, I. Smith, J. Marceau, I. Davey, P. Clark and C. Fell. 2006a. Research Quality Framework. Assessing the quality and impact of research in Australia: The Recommended RQF. Development Advisory Group for the RQF.
- Peacock, J., L. Harvey, M. Barber, P. McFadden, J. Marceau, I. Davey, P. Clark, C. Fell, W. King, T. Enright, D. Schreuder and I. Smith. 2006b. Research Quality Framework. Assessing the quality and impact of research in Australia: Research Impact. Development Advisory Group for the RQF.
- Roberts, G., E. Arthur, M. Barber, R. Batterham, P. Callaghan, I. Chubb, P. Clark, A. Edwards, P. Hoj, R. Milbourne, A. Pettigrew, P. Sheehan and I. Smith. 2005. Research Quality Framework. Assessing the quality and impact of research in Australia: The Preferred Model. Expert Advisory Group for the RQF.
- Spaapen, J. 2005. Judging research on its merits: An advisory report by the Council for the Humanities and the Social Sciences Council. Royal Netherlands Academy of Arts and Sciences.
- Spaapen, J. 2008. Comprehensive evaluation: Balance between research quality and relevance (The Dutch models). *Presentation to Coimbra Group*.
- Spaapen, J., H. Dijkstra and F. Wamelink. 2007. Evaluating Research in Context: A method for comprehensive assessment. Available at http://www.eric-project.nl/nwohome.nsf/pages/NWOA_6V4BXK_Eng as of 22/04/09.
- Trochim, W. 2006. An Introduction to PART. American Evaluation Association. Available at <http://www.eval.org/aea06.williamtrochim.asp> as of 22/04/09.
- VSNU, NWO and KNAW. 2003. Standard Evaluation Protocol 2003–2009 For Public Research Organisations.

- Wooding, S. 2008. *Mapping the impact – Exploring the payback of arthritis research (RB)*. Cambridge, UK: RAND Europe.
- Wooding, S., S. Hanney, M. Buxton and J. Grant. 2005. Payback arising from research funding: evaluation of the Arthritis Research Campaign. *Rheumatology (Oxford)* 44 (9):1145–56.
- Wooding, S., S. R. Hanney, M. J. Buxton and J. Grant. 2004. *The returns from arthritis research: approach, analysis and recommendations. Volume 1*. Cambridge, UK: RAND Europe.
- Wooding, S., E. Nason, T. G. Starkey, S. Hanney, and J. Grant. Forthcoming. *Mapping the impact - Exploring the payback of arthritis research*. Cambridge, UK: RAND Europe.