What's in Good?

Report written by Ethel Méndez

Introduction

Attention to research excellence evaluation has increased in the last few years as governments in England, Australia, and other countries have started exploring ways to allocate research funds on the basis of the quality of research produced. However, what constitutes good quality research or research excellence has long been discussed; not only because of how it may inform funding, but also because of the role that research can play in society. Research helps explain the world around us and can develop, test, and prove new ideas and applications. It allows us to explore the unknown and promises to solve many of the world's problems. Research should be of good quality if it is meant to fulfill that role. But, if high quality research or research excellence is desirable, what do we mean by it? How do we identify research excellence?

This report presents an overview of the literature on research excellence evaluation and exposes some of its gaps. The focus is on what constitutes research excellence and on mechanisms to evaluate it. The literature reveals that there is no single definition, standard, or method for research excellence evaluation. Rather, there are many definitions for both research and excellence, there is no agreement on the quality dimensions that should be used to evaluate research, and there are large debates around the mechanisms used to evaluate research excellence (e.g., peer review and bibliometric analysis). This paper does not answer questions about which definition or approach is better; instead, it presents the range of arguments and ideas found in the literature.

The paper is divided into four main sections: (1) definitions, (2) the 'what' of research evaluation, (3) the 'how' of research evaluation, and (4) main debates in research evaluation. The first section on definitions discusses some of the terms used to describe research, highlighting some of the challenges that discrepancies in terms bring to evaluation. The second section introduces the most common elements of research excellence mentioned in the literature. The third section describes some methods and tools used to evaluate research excellence. While peer review and bibliometrics are at the core of the 'how,' they are not discussed in the third section of the paper. Literature around these two mechanisms is so rich that they have been included in the fourth section, which discusses the main debates discussed in the literature. A sub-section on research impacts as criteria for excellence is also included in the fourth section. The paper concludes with some thoughts on the gaps in the literature and possible areas for further inquiry.

Documents in this literature review and limitations

This report is part of a larger study on research excellence conducted by the International Development Research Centre (IDRC). While the concepts discussed should be useful to researchers and research institutions concerned with how to evaluate the quality of research in any field, the literature review was conducted with a special focus on research for development. This explains why several examples in this paper stem from that field.

The documents included in this paper were selected for detailed review from a larger set of papers generated through searches on databases of peer reviewed journals, Google and Google Scholar, and through targeted requests to librarians, scholars, and via list-serves. Papers were chosen on the basis of a review of their abstracts. Colleagues from IDRC also contributed articles of interest.

The large majority of sources come from scholars in the global North, particularly England, Australia, and the United States, with only three documents from scholars operating in the South. One limitation of this study is that the researcher could only review documents written in English and Spanish.

1. Definitions

Any evaluation process must have a clear definition of what it attempts to evaluate. The case of research excellence evaluation is no different. When talking about evaluating re-search excellence, what do we mean by research? How do distinctions between basic, applied, revolutionary, or interdisciplinary research affect the way we evaluate its quality? What is the distinction between quality research and research excellence? The following paragraphs provide definitions that address those questions.

A. Research quality or research excellence?

There is no agreed term or definition to qualify *good* research. Linda Graham, Robert Tijssen, Jonathan Grant, Philipp-Castian Brutscher, Susan Ella Kirk, Linda Butler, and Steven Wooding argue that quality and excellence, the most commonly used terms to indicate good research, mean different things to different people. Indeed, the literature reveals a few overlaps between the two terms.

A document from the National Center for the Dissemination of Disability Research, cites various authors describing research quality from an almost strictly technical and methodological standpoint. It notes: *"Quality research most commonly refers to the scientific process encompassing all aspects of study design; in particular, it pertains to the judgment regarding the match between the methods and questions, selection of subjects, measurement of outcomes, and protection against systematic bias, non-systematic bias, and inferential error"*(National Center for the Dissemination of Disability Research, 2005, p. 2).

Boaz and Ashby argue that "conceptualizations of research quality need to move beyond a fixation with methodological quality, to address the 'fitness for purpose' of research" (Boaz & Ashby, 2003). They identify dimensions of 'fit' including the "fit of the methods to the aims of the research" and "the fit of the research to the ways in which the findings are likely to be used" (p.12). Writing particularly about research for development, Yule also notes that utility, accessibility, and quality of outputs geared to users are important dimensions of research quality (Yule, 2010, p.1).

Box 1: The UK's Research Assessment Exercise (RAE) and Research Excellence Framwork (REF)

The RAE was the mechanism the Higher Education Funding Council for England used to evaluate research quality. The results from the RAE informed funding decisions for higher learning institutions. RAE was first used in 1986 and was last conducted in 2008. It was a peer-review based system that looked at three dimensions of research: originality, rigour, and significance. Higher learning institutions were ranked according to a five point scale with the highest institutions earning the title of world-leading in terms of originality, significance, and rigour (Higher Education Funding Council of England, 2011).

The RAE was criticized for being unscientific, subjective, expensive, time consuming, prone to bias, and for making comparisons across disciplines that could not be compared due to the different weights they place on quality indicators (Kenna and Berche, 2011). Researchers in universities, the very target for its evaluation, were especially critical of RAE and what it de facto incentivized (i.e., publishing, that too in peer-review journals, as opposed to quality of teaching conducted).

The Research Excellence Framework (REF) was suggested as an alternative to the RAE. The new system, which will be completed in 2014 and will look at research outputs, impact, and research environment, sparked much criticism for its focus on metrics and impact.

At the time this paper was written, the draft criteria for the REF suggested that review panels would decide on indicators and on the use of metrics, but the debate on research impact, how to define it, measure it and whether it should be included at all, continues.

While Yule, Boaz, Ashby, and the OECD consider utility of research an important element of 'quality,' there are others like Grant et al. who take the use or impact dimension as the point of distinction between 'quality' and 'excellence,' where impact is separate from quality but together compose the research excellence framework (Grant, Brutscher, Kirk, Butler, and Wooding, 2010). Their views have informed the new Research Excellence Framework (REF) in England (see Box 1 above).

Robert Tijssen (2003) emphasizes the utility dimension of research excellence and offers three additional elements to distinguish it from research quality. First, he explains that the term 'excellence' implies "surpassing something or someone in some quality" (p. 92), which implies that excellence is a comparative characteristic. Under this approach, excellence can only be determined if compared against other research deemed excellent. Second, research excellence is driven by the "creation of new, high-quality scientific and technical knowledge," which speaks to methodological soundness but also to its originality. Finally, Tijssen flags the

dimension of excellence that relates to publication and commercialization of research products.

Maureen O'Neil, former president of IDRC, suggests a different take on research excellence applicable to the field of research for development. She suggests that "by excellence we may mean 'urgently needed and challenging research' – that which is problem-oriented, multi-disciplinary (preferably comparative) and carried out by teams networking internationally across research sites and policy jurisdictions" (O'Neil, 2002).

Overall, one can argue that the overlap and divergence in the definitions for research excellence and research quality may reflect the belief that the terms can be used interchangeably. However, the fact that some authors like Grant et al. see a clear distinction indicates the need to clarify and draw boundaries. The distinction based on research impact is helpful and seems to be understood by people who are familiar with England's REF, but it is not widely agreed upon.

This review is about research excellence, which is interpreted to include broader dimensions of research that go beyond its methodological or scientific rigour. However, since the distinction between 'quality' and 'excellence' is not explicit nor acknowledged in many of the papers reviewed, and since the intent is to present the range of views found in the literature, this report cannot commit to one precise term. Rather, the author assumes a broad interpretation of quality and uses the terms 'quality' and 'excellence' interchangeably.

B. Basic and applied research

"When we are identifying assessment criteria as a basis for accountability we need to recognise that this depends upon what is taken to be the proper task of research"(Hammersley, 2008, p. 754).

The main distinction between basic and applied research is that the former is conducted with the intent of advancing knowledge and theoretical understandings while the latter attempts to generate findings that can solve a problem (Coryn, forthcoming 2013, p.9). Stewart Donaldson and Christina Christie add that the origin of the questions of the study and the settings in which the work is conducted also differ between applied and basic research. Applied research, they note, "is problem based or solution oriented, and conducted in 'real world' settings as opposed to highly controlled, traditional scientific laboratories" (Donaldson, 2009, pp. 2-3). Donald Stokes also introduces the concept of use-inspired research, which is research that, like basic research, advances knowledge and scientific understanding, and like applied research, has potential practical applications (Arnold, 2008, p. 2189).

These definitions are important in evaluation of research excellence because the elements to be considered in determining the quality of research will depend on the purpose it intends to serve. John Furlong and Alis Oancea, for example, argue that the criteria to evaluate the quali-

ty of applied research must integrate dimensions of research application and use in addition to the knowledge generation dimension that would be evaluated in basic research (Furlong & Oancea, 2005, p. 8).

C. Revolutionary research and emerging disciplines

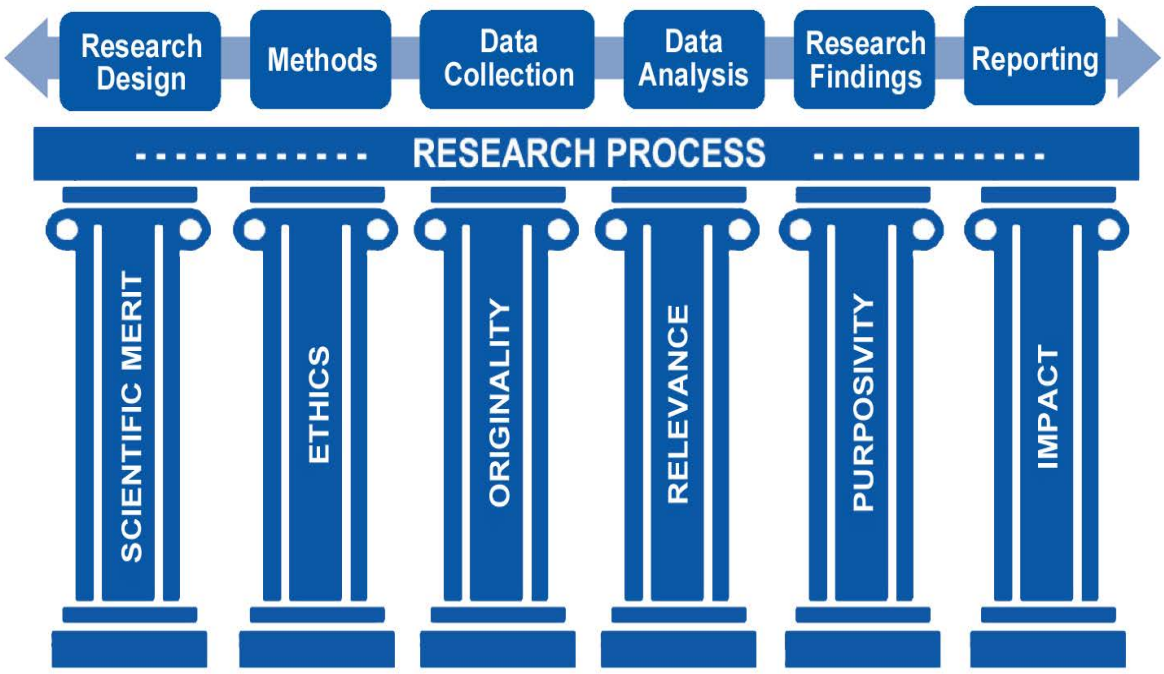
"It is with no inconsiderable degree of reluctance that I decline the offer of any Paper from you. I think, however, you will upon reconsideration of the subject be of opinion that I have no other alternative. The subjects you propose for a series of Mathematical and Metaphysical Essays are so very profound, that there is perhaps not a single subscriber to our Journal who could follow them" (Rodrik, 2011).

Danni Rodrick includes these lines in his blog entry *'A rejection letter I would like to receive from a journal one day.'* It is part of a letter that Charles Babbage, known as 'the father of a computer,' received in 1821 from The Edinburgh Journal of Science as cited in James Gleick's *The Information: A History, a Theory, a Flood.* While Rodrik's comment can be taken with some humour, it raises an important issue about the evaluation of revolutionary research. "Revolutionary research can be understood as research that "aims to bridge either a relatively big gap in the theoretical sense (involving many steps of deductive reasoning) or a relatively big gap in the sense of experimental data collection (meaning the development of radically new techniques)" (Andras, 2011, p. 94).

Peter Andras builds on Kuhn's idea of revolutionary science, which generally involves a change in basic assumptions or a paradigm shift (Kuhn, 1962). Since revolutionary science does not have the rich theoretical tradition that normal science has, new findings in revolutionary or discovery science run the risk of dismissal for being ahead of their time. Such findings often have a hard time breaking into the scientific community, which delays the recognition of the discovery, its publication, and application. This delayed recognition, in turn, affects research evaluation, particularly when excellence is tied to publication rates. Andras explains, *'In terms of publishing research results in high ranking journals, the normal science work has much better chances, as it is more likely to be readily accepted by scientific public opinion than a potentially controversial revolutionary science result'* (Andras, 2011, p. 97).

Similarly, the Organization for Economic Cooperation and Development (OECD) notes that "there are special problems involved in evaluating […] emerging disciplines, which are often inappropriately appraised by current bibliometric countings, as well as by more qualitative measures based on peer judgments" (Organisation for Economic Co-operation and Development, 1997, p. 9). In peer review processes, revolutionary research faces the challenge of identifying peers who are qualified to evaluate its quality (Wooding & Grant, 2003, p. 25).

Diagram 1: Common Conceptual Elements of Research Excellence in the Research Process



D. Interdisciplinary research

Veronica Boix Mansilla, Irwin Feller, and Howard Gardner describe interdisciplinary research as "a form of inquiry that integrates knowledge and modes of thinking from two or more disciplines or established fields of study to produce a cognitive or practical advancement" (Boix Mansilla, Feller, & Gardner, 2006, p. 70). Craig Stephen and Ibrahim Daibes add that while the production of new and cutting edge knowledge is important, in interdisciplinary fields like global health research the breadth of the research is at least as important as its depth (if not more so) (Stephen & Daibes, 2010, p.4). Such considerations must be taken into account when evaluating the quality of interdisciplinary research.

Evaluation of the quality of research that integrates various disciplines is also problematic because disciplines have different standards of excellence that do not necessarily match and sometimes conflict. For example, Annette Boaz and Deborah Ashby explain that positivist and constructivist epistemological paradigms underlying disciplines can make evaluation of interdisciplinary research problematic: *"methodological debates in the natural sciences focus on the quest for 'truth' and the elimination of bias. In the social sciences the existence of objective truth is often contested, while bias is often an accepted dimension of knowledge, to be acknowledged rather than eliminated"* (Boaz & Ashby, 2003, p. 9).

These differing views on methods, objectivity, and truth result in differing views on methods-based standards of excellence, which may come into conflict when looking at interdisciplinary research. Clinical Randomized Control Trials (RCT), for example, are often consid-

ered the gold standard for excellence in healthcare research, but their use in the social sciences, particularly in the interdisciplinary fields that characterize international development like governance, information communication technologies, and others, are increasingly being contested.

Finally, selecting adequate research evaluators and managing their concurred expertise during the evaluation of interdisciplinary research is daunting. Like revolutionary research, the problem is that there are few peers who have the right combination of interdisciplinary knowledge to evaluate certain types of research. As a result, evaluation of the quality of interdisciplinary research is often conducted along discipline lines which, as mentioned above, can be problematic (Boix Mansilla, Feller, & Gardner, 2006, p. 70). (Boix et al. propose having strategically constructed review panels that embody multiple disciplinary perspec-tives as well as involving 'interpreters' who can bridge the epistemic gaps among experts. They also suggest participation of university officials or program officers who can identify innovation and overcome the conservative tendencies of review panels (Boix Mansilla, Feller, & Gardner, 2006, pp. 70-71).)

2. The 'what?' of evaluating research: Excellence frameworks

This report has already discussed some of the basic, often contentious definitions of research evaluation. However, putting those discussions aside, what should be the criteria for research excellence?

Virtually every research or methods text book or handbook includes one or various sets of quality dimensions. The literature reviewed for this paper cites at least 30 sets of criteria which vary in length, detail, and approach. Looking across those 30 sets revealed recurring conceptual elements and specific criteria used in evaluating research excellence. The conceptual elements - purposivity, relevance, originality, scientific merit, ethics, and impact - are discussed in the following section and highlighted in Diagram 1 (see page 6).

The more specific criteria that unfold from these conceptual elements and underscore the research process can be found in Annex 1.

A. Conceptual elements of research excellence

Purposivity. Research must have a clear purpose and research questions should be informed by it (Aagaard-Hansen & Svedin, 2009). Aagaard-Hansen & Svedin connect purposivity to having a well-formulated problem, well-defined terminology, consistency, and demonstrating knowledge of relevant literature (2009). Supporters of action research, for example, indicate that the purpose is to generate transformative processes. The research design – questions, approach, methods and tools – are then selected to respond to that purpose (Anderson and Herr, 1999 as cited in Groundwater-Smith & Mockler, 2007).

Box 2. Ranking research quality criteria

A report commissioned by the Lavenham Suffolk Social Policy Association and the Joint Universities Council Social Policy Committee in England summarized the findings of several workshops where over 250 researchers and research managers discussed and ranked 35 research quality criteria. The findings included:

- The highest ranking was assigned to 'accessibility of research output to appropriate audience' and 'design that is adequate to the questions.' Publication criteria ranked lowest.
- Among methods, in-depth interviews were associated with the highest quality while RCTs, which are usually perceived as the gold standard in research, were associated with lower quality.
- In quantitative research, validity and reliability ranked highest. Other criteria included: replicability, generalizability, robustness, transparency of methodology, congruence, clearly specified research questions, and appropriate hypothesis testing.
- For qualitative research, validity was important and reliability, replicability, and generalizability were not considered as important. Credibility and confirmability were considered somewhat important as were explicitness and transparency, relevance and involvement of users, and reflexivity.
- For mixed methods research, most researchers did not favor using the same criteria for both quantitative and qualitative elements.
- Participants suggested additional criteria that included relevance to research questions, transparency, need for integration of mixed methods findings, and rationale for mixed methods (Becker, Bryman, & Sempik, 2006).

Relevance. The quality criteria from the Association of Universities in the Netherlands looks at relevance from two standpoints:

• scientific relevance or significance for development of the discipline; and
• societal relevance in terms of societal and technological impacts (Organisation for Economic Co-operation and Development, 1997).

Boaz & Ashby also speak of relevance in terms of research that is relevant for policy and practice: "the extent to which the research addresses the needs of key stakeholders is an important dimension of quality" (p.12). Criteria from RAND Corporation adds that research "should be compelling, useful, and relevant to stakeholders and decision makers" (RAND Corporation, 2011). (The ideas of relevance, impact and purposivity are closely related but presented separately in the literature, hence why they are presented here separately.)

Originality. In a study conducted by Saul Becker, Alan Bryman, and Joe Sempik, a group of 250 researchers and research managers from England agreed that "originality revolved

around viewing existing issues, ideas, or data in a new or different way, more so than generating new data or novel methods. Originality involved, also, the development of new theoretical and practical insights and concepts" (Becker, Bryman, & Sempik, 2006, p. 12). Wooding & Grant also note that "defining the research agenda by framing new research questions and advancing a field into new areas" is one of the main characteristics of high quality research (Wooding & Grant, 2003, p. 14).

Scientific merit. Quality criteria include elements of soundness or rigour of the research methods, different forms of validity (terms varied according to the type of research), and logical interpretations of data. Several of these components are presented in the specific criteria in Annex 1. Some authors noted criteria specific to certain methods such as random and concealed assignment to treatment groups in RCTs (Boaz & Ashby, 2003) and capacity to act free of psychological and organizational pathologies in cooperative inquiry (Martí & Villasante, 2009).

Ethics. Susan Groundwater-Smith and Nicole Mockler advance the idea that "ethics are the primary criteria for quality in practitioner research" (2007, p. 204). They view ethics as "an orientation to research practice that is deeply embedded in those working in the field in a substantive and engaged way" and suggest ethical guidelines for practitioner research (p.206):

- That it should observe ethical protocols and processes.
- That it should be transparent in its processes.
- That it should be collaborative in its nature.
- That it should be transformative in its intent and action.
- That it should be able to justify itself to its community of practice.

Impact. Supporters of research impact as an element of quality hold that research should be useful in order to be high quality. Research evaluation should, therefore, look beyond the research outputs and consider outcomes, influence or changes caused by the research. There is much debate around how to define and measure research impacts and even whether it should be a dimension of research excellence. The primary arguments in this debate are discussed later in this report.

B. Weighting research excellence criteria

Different authors placed emphasis on different quality criteria (see Box 2 on page 8 for a study on ranking of quality criteria). Some of the variation in emphasis can be explained by different units of analysis. (The literature identified the following units of evaluation: outputs; individuals; research teams; laboratories and institutions such as universities; scientific discipline; government programs and funding agencies; a country's entire research base, etc. (OECD, 1997; Rons, De Bruyn, & Cornelis, 2008; Hammersley, 2008). Laudel and Glasel add that it is im-

portant to define what counts as outputs if output-based indicators will be used in the evaluation (Laudel & Glaser, 2006).) For example, the REF - where the unit of analysis is higher level education institutions in England - considers elements of the 'research environment' like research doctoral degrees awarded, research income, and research income in-kind. These elements would not make sense in other evaluation processes where the unit of analysis may be research outputs. However, in the case of REF, looking strictly at research outputs could create incentives for researchers to focus on publication, therefore jeopardizing the teaching, mentoring, and other scholarly activities that are required at an institutional level.

3. The 'how' of research excellence evaluation

The most commonly used mechanisms to evaluate research are peer or expert review and bibliometric analysis. Both have been heavily criticized and are discussed in detail in the following section of this paper. However, despite the predominance of peer review and bibliometric analysis, several people are experimenting and promoting other mechanisms to evaluate quality. This section identifies some of those alternative mechanisms, some of which combine elements of peer review or metrics with other tools and approaches.

Historical ratings: Defined by Wooding and Grant (2003) "as a system in which ratings of groups/ departments/ universities are determined entirely by their performance in the past" (p. 20). The benefits of this approach are that it is inexpensive, transparent, simple, and focuses on track record rather than ambition. The disadvantages include that it doesn't cope well with change, may perpetuate inequalities, encourages complacency, does not motivate improvement, is purely retrospective, and lacks credibility (Wooding & Grant, 2003, p. 24).

Self-assessment: Wooding and Grant describe it as a system where "institutions, departments or individuals assess themselves" (p.20). It has been criticized by researchers and research administrators in England for lacking credibility, being burdensome, subjective, and prone to bias, allowing for gaming and overrating, not being comparable, and being too inward looking. Benefits included simplicity and low cost, flexibility and sensitivity to local conditions, promoting learning, reflection, and responsibility, and requiring the engagement of researchers in the process (Wooding & Grant, 2003, p. 27).

Computerized semantic analysis: The European Union Framework 7 EERQI project is exploring the possibility of using computerized semantic analysis to identify key features in documents as a basis for quality evaluation (Bridges, 2009, p.507). Bridges indicates that, while the intention is to identify key features in documents, the idea of a computerized system is inadequate given that such an analysis does not offer the benefits of an informed, experienced perspective.

Box 3. Mechanisms to Evaluate Research Impacts

In a review of international practices, Grant et al. explored four existing research impact evaluation methods that could inform England's REF. They looked at the Australian Research Quality and Accessibility Framework (RQF) which is a case study approach, the UK RAND/Arthritis Research Campaign Impact scoring system (RAISS) which is an indicator based approach, the US Program Assessment Rating Tool (PART) which is a self-evaluation approach, and the Dutch Evaluating Research in Context (ERiC), which is a an approach that mixes some of the elements of the other systems.

The report concludes that all systems were burdensome and generated undesirable incentives, yet the case study approach of the Australian RQF showed the most promise for the REF (Grant et al., 2010, p. 69).

Benchmarking (metrics-based system): The International Monetary Fund (IMF) conducted an evaluation of the quality of their research in 2011. The study was based on citation and publication counts and was conducted across eight similar policy institutions, ranking research quality against one another. The study acknowledges the shortcomings of benchmarking as an "approach [that] is still subject to possible biases related to differential scales, heterogeneity, and the varying missions of the eight policy institutions" (Aizenman, Edison, Leony, & Sun, 2011, p. p 19). (The study concludes that self-citations are high within the IMF, higher than the other studied institutions, which could skew perceptions on impact. For its size, the IMF has relatively few publications.)

Tijssen (2003) also explores the use of scoreboards, which he defines as a "mode of comparative analysis based on structured collections of relevant quantitative data of various aspects of research excellence […] designed specifically to systematically open up [sic] the range of quantitative indicators for further examination and comparison – both for analytical and representational purposes" (p. 96).

Case studies: In a study on the history of research evaluation, RAND Corporation (2009) found that case study approaches have dominated the research field in evaluating research contributions to society (Marjanovic, Hanney, & Wooding, 2009, p. 17). Grant et al. describe the case study approach as one that relies on evidence-based impact statements that include quantitative and qualitative information. The evaluation then relies on expert opinion. Australia's Research Quality Framework was designed as a case study based impact evaluation system but was not implemented because of a change in the Australian government (Grant et al., 2010, pp. 7-20).

Indicators: Grant et al. examine the RAND/Arthritis Research Campaign (ARC) impact scoring system, which is an indicator-based system consisting of 187 yes and no questions that aim to capture information on research grants awarded by the ARC. The process captures impacts

that have occurred and emphasizes wider impacts like improving research capacity. This approach has the advantage of a short completion time, but the yes or no answers do not allow for diversity of responses or other explanations that may explain findings (Grant , 2010, pp. 21-31).

Mixed approaches: The Dutch Evaluating Research in Context (ERiC) is a mixed methods approach that integrates self-assessment, data gathering based on areas of focus, including mapping of outputs, and consultation with key stakeholders through interviews and surveys (a form of peer review). The workload involved in this type of evaluation is heavy but the multiple data sources and steps involved result in a comprehensive understanding of quality and provide learning for the institutions (Grant, Brutscher, Kirk, Butler, & Wooding, 2010, pp. 47-58).

Deliberation and consensus: The National Center for the Dissemination of Disability Research (2005) comments on the importance of deliberation and consensus as a way to determine research quality: "Consensus among a community of scholars is one of the most respected means of quality assessment. Strategies for reaching consensus include position statements, conferences, the peer review process, and systematic review."

Bridges (2009) and Stephen and Daibes (2010) also note the importance of consensus and suggest a form of peer review that engages multiple stakeholders and disciplines in deciding criteria (if any). In this model, the review and the decisions about quality is potentially more dynamic than a straightforward expert review of documents process that has no interaction among experts. Drawing from Martha Nussbaum's *Love's Knowledge*, Bridges notes that flexibility, responsiveness, and openness in the process are essential for good deliberation.

4. Key debates in research excellence evaluation

Pressure on research funds and the focus on results-based management and efficient use of resources, particularly within the education sector in England and Australia, have led to heightened attention on the quality of research produced and, consequently, on how to evaluate it. In that context, a number of debates have emerged around the evaluation of research excellence. This section discusses three such debates, namely (1) research impacts as a dimension of excellence, (2) peer or expert review, and (3) bibliometrics as mechanisms for research excellence evaluation.

A. The impact debate

While discussions about evaluation of research impacts date to the 1960s (Marjanovic, Hanney, & Wooding, 2009), England's decision to change the RAE (which evaluated the originality, rigour, and significance of research) to the REF (which incorporates research impact) re-ignited the debate in England and abroad. The issue of research impacts as an element of research

quality evaluation can be summarized into three main concerns: What is impact? How should impact be measured? Should impact be a dimension of research evaluation?

What is impact? In an extensive literature review conducted in 2003 that focused on models for research impact, Sandra Nutley, Janie Percy-Smith, and William Solesbury found little theorizing or discussion on definitions for research impact. They did, however, identify the following forms of research impact (2003, p.11):

- changes in access to research
- changes in the extent to which research is considered, referred to or read
- citation in documents
- changes in knowledge and understanding
- changes in attitudes and beliefs; and
- changes in behaviour.

Drawing on Huberman, they make a distinction between the conceptual use of research which, they note "brings about changes in levels of understanding, knowledge, and attitude" and the *instrumental* use of research which "results in changes in practice and policy making" (Nutley, Percy-Smith, & Solesbury, 2003, p. 11). Since their study, the debate has intensified, raising new definitions and ways of looking at research impacts.

Funded by the Higher Education Funding Council of England (HEFCE), the London School of Economics (LSE) launched a multi-year project that "aims to demonstrate how academic research in the social sciences achieves public policy impacts, contributes to economic prosperity, and informs public understanding of policy issues and economic and social changes" (London School of Economics, 2011). In a handbook developed to help researchers maximize the impact of their work, the group defines research impact as "an occasion of influence and hence it is not the same thing as a change in outputs or activities as a result of that influence, still less a change in social outcomes" (LSE Public Policy Group, 2011, p. 21). The document categorizes research impacts into:

- academic impacts which are described as instances when research influences actors in academia or universities. These are measured by citations in the work of other academics; and
- external impacts, described as instances when research influences actors like business, government, or civil society, outside of higher education. External impacts are measured by references in the "trade press or in government documents or by coverage in the mass media" (LSE Public Policy Group, 2011, p. 5).

Eveliina Saari and Katri Kallio also provide a definition for research impact. Drawing on Donovan's (2008) definition of research - "adds to the social, economic, natural, and cultural capital of the nation" - they suggest that high quality research must have an impact at three levels:

accumulating and combining research knowledge; solving the client's problems; and addressing a societal question (Saari & Kallio, 2011, p. 232).

How to measure impact? Questions about measuring research impact go beyond the difficulties of not having a clear definition. Yates explains that, in Australia, "impact in education research is currently well understood in terms of research designed to speak to policy makers and to system questions" (Yates, 2005, p. 395). In that context, where impact is understood beyond academic impact, the questions focus on how to measure impact and on who should do it. Values, Yates notes, play a role in determining good and bad impact. For example, it is widely agreed that Hitler had a large negative impact in society (Yates, 2005). However, not all situations can be judged that easily and there may be conflicting views on whether an impact is positive or not. This speaks to the challenges posed by the paradigms underlying evaluation, an issue that Yates labels the value diversity problem (Yates, 2005, pp. 398, 401).

The debate on measurement includes, on the one hand, those who believe that bibliometrics are good indicators of impact; and, on the other hand, those who seek other approaches for measuring impact. For example, in terms of bibliometrics, R. Tatavarti, N. Sridevi and D.P. Kothari describe the Impacts Factor (IF) as "a measure of the citations to refereed journals in science, humanities, and engineering… frequently used as a tool to gauge the relative importance of a research journal within its field" (2010, p. 1015). They hold that the IF is among the most robust quantitative measures of research quality. As mentioned earlier, the evaluation of IMF research conducted in 2011 relied on bibliometric analysis, tracing citations as a proxy for short term impact and publications as a proxy for long term impact (Aizenman, Edison, Leony, & Sun, 2011). On the other side of the spectrum there are studies, like the one described in Box 3 on page 11, that highlight other approaches to identifying research impacts.

Despite the arguments around bibliometrics, the main problem with measuring research impacts is attribution. Hammersley states quite assertively, that "returns on research are not calculable" (Hammersley, 2008). Citing Carol Weiss's idea of knowledge creep, Furlong and Oancea also note that it can take over 20 years to bridge the knowledge-policy gap; thus, measuring impact in the short term is problematic (Furlong & Oancea, 2005). (Carol Weiss's idea of 'knowledge creep' is that research often does not have a direct impact or influence on policy or practice. Rather, ideas stemming from research slowly 'creep' into policy and practice settings, changing assumptions and raising questions over a long period of time.) Frances Seymour, Director General of Center for International Forestry Research (CIFOR) blogs about the same issue in a post entitled "Does the pressure for impact compromise research?" She writes, *"contributing to this set of challenges is the difficulty of attributing impact, especially when it comes to policy research, and research designed to influence natural resources management (NRM) rather than (for example) to increase the yield of a particular grain crop"* (Seymour, 2011).

While not explicit in their writing, these authors allude to the difficulty of ascribing causation in environments that are complex, where there are different systems at play, and where there are multiple causes and pathways to achieving impact.

Should impact be considered in research quality evaluation? The discussions on defining and measuring impacts are based on the assumption that impacts should be considered in the evaluation of research excellence. That assumption does not go uncontested. Hammersley (2008), for example, criticizes the emergence of an 'investment model' in research. That is, a model where "a return is demanded on individual [research] projects, and evidence showing the payoff and efficiency of knowledge production across the sector is demanded" (p. 753). That model opposes what he calls a state patronage model where "the production of knowledge was treated as beneficial in itself" (p. 753). Incalculable returns on investment, negative effects on academic freedom, and delayed effects on policy are among the reasons why he opposes the model, but he also contends that the "investment model may also be at odds with any commitment to research informing public discussion of policy issues, since it frames inquiry within assumptions about predictable payoff" (p. 753).

Writing from the education field in Australia, Yates makes similar arguments and adds that, given the ties that evaluation processes have to funding, the focus on impacts may not be adequate as it may create incentives for research institutions to focus on topics where they can have the fastest results, compromising academic freedom.

B. The peer review debate

"Expert Review (including Peer Review): A system in which experts make a professional judgment on the performance of individuals or groups, over a specified cycle, and/or their likely performance in the future. The groups could be research groups, departments or consortia. Assessment may be under-taken entirely by peers or may incorporate other experts such as representatives of user groups, lay people, and financial experts" (Wooding & Grant, 2003, p. 20).

Peer review is the most cited method for evaluating research excellence (Jongbloed et al, 2000; CRE, 2000 as cited in Rons, De Bruyn, & Cornelis, 2008); however it is also subject to heated debate about its efficiency and effectiveness. The sections that follow describe some of the main concerns about peer review.

Unhealthy competition: The literature argues that peer review triggers an unhealthy competition and rivalry among peers, sometimes even motivating scholars to delay publication of other scholars' work in order to publish themselves first (Rowland, 2002; Roebber & Schultz, 2011). Petit-Zeman adds that it is a 'competitor review' that researchers use to "shoot down their rivals" (2003, para. 8). Other authors also note issues of bias refereeing and plagiarism.

Subjectivity: Tijssen notes that peer review "relies on qualitative judgments of a few leading experts, thus introducing a restriction of scope and risking biases due to subjectivity" (2003, p. 101). Further, review panels in university settings have been criticized as inconsistent. According to Tijssen, subjectivity underlies all research evaluation processes, even bibliometrics. In order to receive citation counts, for example, a paper goes through a review process where peers determine whether it is worth publishing. It is not until after the review – where the subjectivity is injected - that citations counts and other bibliometrics can be tracked.

Expertise of panels: The literature expressed concern over who evaluates, what qualifications and capacities theses evaluators must have, and who determines who is qualified to be a peer reviewer/evaluator. Similar to the value diversity problem in evaluation impact, philosophical or methodological biases are also a concern (Yates, 2005, p. 397).

Resource intensive: Lawrence O'Gorman (2008, p.3) criticizes the lengthy and time-consuming process involved in peer review and argues that it discourages researchers from accepting papers for review. Peer review is also expensive. As mentioned in Box 1 on page 3, this was one of the criticisms of the UK's RAE, which led to the proposal of a metrics-based system. Mark Ware explores this issue and pinpoints creative market-based alternatives to facilitate peer review, including the possibility of a virtual currency where researchers accumulate credits based on the papers they review (2011, p.48).

Conservative: As noted in an earlier section, new and revolutionary fields face the challenge of finding peers with qualified expertise to evaluate their work. O'Gorman raises the issue that small fields are often dominated by one way of thinking, therefore making the review process even harder for those trying to break ground (2008, p. 5).

Unclear purpose: Peer review mechanisms are widespread, but their purpose is often unclear. Ware notes: *"To start with, it is difficult to decide how to improve the [peer review] system without agreement on its goals (Ware, 2006). Is it to select the best papers to publish in a journal? To minimize (if it cannot eliminate) fraud and other misconduct? To improve the quality of papers published? To improve the quality of research? To act as a filter, by rejecting bad work, or by deciding where a paper is published rather than whether it is published?"* (p.33).

This is an important concern because the validity of the review depends on what the review is intended for. Without a clear purpose, one can misinterpret the results.

Other criticism described in the literature concerned peer review's inability to detect fraud, lack of transparency in selection of panels, failure to guarantee high methodological standards, slow and incompetent process, and failure to look at the relevance of the research to policy-making (Boaz & Ashby, 2003; Yates, 2005).

Despite the challenges raised, many researchers agree that peer review is the key element of research evaluation, for there is no other way to determine the quality of research if not by reading it and offering an informed opinion (Becker, Bryman, & Sempik, 2006, p. 19). For instance, a report commissioned by the Joint Funding Bodies' Review of Research Assessment in England described the findings of nine workshops where 142 researchers and research managers from higher education institutions gathered to explore research quality and attitudes towards different evaluation mechanisms. Participants were asked to rank four types of research evaluation systems: metrics-based, historical ratings, self-assessment, and expert review. The findings showed that expert review had the "least good features and most bad features" in comparison to the other systems (Wooding & Grant, 2003, p. 25). However, when participants were grouped and asked to design the ideal research evaluation system, 22 out of the 29 groups based their system on a peer or expert review process.

Acknowledgement of the flaws of peer review has led many institutions and scholars to experiment with different types of review to reduce bias and increase efficiency. Ware lists a few: rating reviews and giving feedback to reviewers, providing checklists or templates for reviewers, training reviewers, and offering rewards to reviewers (2011, p.34). Some new models for peer review are also being tested, such as author pays model (Rowland, 2002), post-publication peer review, open review, and cascade review (Ware, 2011). (On the author pay model: the author warns that such an arrangement would place some researchers, including many from the global South at a disadvantage. In cascade review, the rejected author is asked whether "they wish to have their paper re-submitted with its previous reviewer reports attached, thereby reducing the amount of reviewing required by the next journal" (Ware, 2011, p. 35).) Technology has also helped improve the process as the internet has facilitated the exchange and tracking of articles, reviews and reviewers. Yet, multiple authors say that technology offers many more opportunities to improve the system.

Overall, in referring to the debate about peer review, Ware notes, that "instead of a collapse [… we see a wide range of debate, experimentation, and innovation across a growing number of journals. Far from being in crisis, it could be argued that peer review has never looked more vibrant in its growing diversity" (2011, p. 49).

C. The metrics debate

The use of metrics for research excellence evaluation is highly contested. According to Andras, metrics include: *"publication count in defined list of venues (journals, conference proceedings) by individuals, departments, or universities, the citation count of these publications over a defined period of time, indexes calculated using these counts (e.g., h-index), metrics derived from citations and authorships graphs, and market share measures"* (Moed, 2005; Moed et al, 2004; HEFCE, 2008a referenced in Andras, 2011, p. 90).

Metrics are regarded as cheaper, less burdensome, and more objective and transparent alternatives to peer review (Wooding & Grant, 2003, pp. 22-23). Coryn explains that citation counts are perceived as indicators of performance in research evaluation based on the assumption that "highly cited research can be considered meritorious or significant, since the extent to which research is used (i.e., cited) is a measure of its contribution to knowledge" (2006, p. 115). However, even supporters accept that metrics have many limitations and much needs to improve in order to overcome their weaknesses (Andras, 2011, p. 103). The following paragraphs describe some of the main issues around metrics.

Validity: Perhaps the most critical argument against metrics is that they are not true indicators of quality, but proxies for it. Bridges makes a distinction between Type A and Type B indicators where Type A are measures of the intrinsic quality and Type B are proxies for quality. In the case of a diamond, he argues, a Type A indicator could be the clarity of the diamond while the price would be a Type B indicator. He contends that evaluation based on type B indicators can distort evaluation; in his diamond example, the price of the diamond may vary for reasons other than its quality. Similarly, he sees metrics as proxies for quality in research. He states: *"Any quality assessment system needs to ensure that this assessment is based on Type A indicators, on characteristics which are intrinsic to the quality of the work, rather than indicators which are not intrinsically about quality, the use of which will distort academic practice but do nothing to improve quality"* (2009, p. 511).

Boaz and Ashby raise the same concern when they say that "it is a faulty assumption that all research that is published in journals or cited by others is accurate, reliable, valid, free of bias, non-fraudulent, or of sufficient quality" (2003 as cited in National Center for the Dissemination of Disability Research, 2005, p. 2). After all, citation counts and other bibliometrics rely on peer review and other subjective decisions; for example, a paper undergoes a peer review process before being accepted into a journal, conference, or other 'venues' from which bibliometrics derive. Journal decisions to publish consider other factors outside the quality of the paper itself, such as the reputation of the author or the popularity of the topic, among others. Ware notes that "it is also the case that almost any genuine academic transcript, however weak, can find a journal to publish it if the author is persistent enough" (2011, p. 29). Bibliometric analysis, therefore, can become a measure of quantity, which can be a reflection of the researcher's commitment to publish or of a journal with high acceptance rate, rather than an indication of the quality of the research (COSEPUP, 1999 as cited in National Center for the Dissemination of Disability Research, 2005, p. 2).

An example of how bibliometrics can provide an erroneous indication of quality comes from clinical drug trials. Ernest House (2008) argues that drug companies have such control over the industry that they can manipulate publishing to suit their needs. House notes that companies are three times more likely to put forward for publication positive – albeit sometimes incom-

plete – drug results and may even write reports for the researchers to make studies sound more favorable (p. 418). When publishing is considered an indication of research excellence, a piece of research may be considered high quality when, in fact, publication and bibliometrics may only reflect the ability of a particular drug company to game the system. House notes that journal editors, like Jeffrey Drazen from the *New England Journal of Medicine*, now understand the problems with publishing clinical trials and are proposing new standards for publication that include: full data disclosure, transparency about financial interests, submission of original designs (to note changes and modifications), and signed statements from authors indicating that they have, indeed, written the report (p. 423).

Further, referring to research in education, Rons, De Bruyn & Cornelis criticize the narrowness and incompleteness of bibliometrics and the dangers when funding is tied to evaluation. They hold that bibliometrics ignore aspects of quality such as, *"[t]he training aspects and international embedding of your researchers, the research culture in which students are immersed, the impact of research on society, economy, culture, government policy, development aid, the scientific literacy of citizens and the non-profit sector"* (2008, p. 46).

Negative citations: Bridges (2009) and a document from the Higher Education Funding Council of England (2011) mention concerns about negative citations, that is, papers that are cited because of their poor quality. While one can argue that these papers may contribute to the knowledge base, evaluation that aims to identify good quality research may be misinformed if it relies on metrics of papers that are cited because they are heavily criticized due to their poor quality. Similarly, writing about download counts as metrics, Bridges (2009) reminds us that a paper can be deemed of bad quality after it is downloaded. Negative citations, like other issues mentioned above, question the validity of metrics as indicators of quality.

Negative incentives: Sastry and Bekhradnia (2006), Bridges (2009), and others have also expressed concern that bibliometrics generate incentives for researchers to focus on publication or to do and conduct research in a way that may increase their chances of publication. In describing publication patterns in South Africa, Robert Tijssen, Johann Mouton, Thed N van Leeuwen, and Nelius Boshoff note that "procedures for evaluting success in universities in developing countries usually measure achievement (and thus allocate rewards) in terms of success in publishing research results in international peer-reviewed mainstream publications" (2006, p. 171). However, the topics of interest in South African policy may not be interesting to journals based in the global North. This places researchers at odds: should they follow the incentives and gain 'rewards' or should they risk losing the rewards and stick to the research agendas that matter in their contexts?

Spurious objectivity: Objectivity is often cited as one of the benefits of metrics based systems. However, while the calculations may be indeed objective, the selection of information consi-

dered in the computation may not. Who makes the decision about what goes into the computations also raises concerns (Wooding & Grant, 2003, p. 23).

The literature notes that journals and other venues that are the basis for metrics are biased in favour of English language research outputs (Yule, 2010; Liang, Wu, & Li, 2001; Bridges, 2009). Literature about publication in China and South Africa indicates that the majority of their publishing occurs in local journals (Liang, Wu, & Li, 2001; Tijssen, Mouton, van Leeuwen, & Boshoff, 2006). As mentioned, the case of South Africa reveals that even when language is not a barrier, the topics, methods, and presentation of the work of South African researchers may not match the requirements or interests of Northern led and/or based journals. This reduces the opportunities for publishing in what are considered high quality international venues and, along with them, the number of citation counts southern researchers receive (Tijssen, Mouton, van Leeuwen, & Boshoff, 2006, p. 172). Relying on metrics for quality evaluation in this case may create the erroneous impression that southern research is of poor quality when, in fact, the lower citation counts may be an indication of a different research agenda.

Fields of research: Coryn (2006), Bridges (2009), Kenna & Berche (2011), and others raise concern about the appropriateness and comparability of metrics between fields and disciplines. Bibliometric data varies across disciplines. A document from the OECD also points to the fact that bibliometrics don't take into account grey literature which is "of cardinal importance for interdisciplinary work as well as for innovative developments" (1997, p.9).

Graham (2008) also notes that researchers in the arts, humanities, and social sciences in Australia rejected the use of Thomson ISI impact factors as proxies for journal ranking because they found them inappropriate for their work. Writing also about the arts, Coryn adds, *"Publication of a work may entail hearing it, viewing it, reading it, or experiencing it in other ways, such as through a performance on a stage or in some other public forum; a narrow view of evaluating publication in the creative arts in written terms alone creates anomalies whereby a painter's paper about their own exhibited painting counts as a publication but the painting does not, and a critical paper on musical composition counts as publication while a performance of the composition, and even the composition itself, do not count as publication and therefore cannot be cited"* (Strand, 1998 as referenced in Coryn C. L., The Use and Abuse of Citations as Indicators of Research Quality, 2006, p. 117).

Innovations and new researchers: As noted earlier, metrics are particularly troublesome when dealing with new fields. Andras, who supports the use of metrics, notes: "In terms of publishing research results in high ranking journals, the normal science work has much better chances, as it is more likely to be readily accepted by scientific public opinion than a potentially controversial revolutionary science result" (2011).

Discussions around the RAE and REF in England have also highlighted the difficulties that young researchers face in publishing since journals may be more willing to include a piece

from a well-known author than one from a new researcher, in spite of its quality. Metrics, therefore, tend to favour more experienced researchers over younger ones.

As a whole, bibliometrics hold some promise. Wooding et al. note that there are ongoing discussions on how to create metrics that are more field-specific and that consider research produced in other contexts, in other languages, and by new researchers. However, there is a strong position among researchers that bibliometrics should not and cannot replace expert opinion. Where appropriate, they can be used as a tool to complement evaluations of research excellence (van Raan, 1993; OECD, 1997; David, 2008; Tatavarti, Sridevi, & Kothari, 2010). Others like Becker, Bryamn & Sempik hold that the quality of research can only be judged "by people involved in the old art of reading" (2006, p. 19).

Conclusions and issues for further discussion

Similar to the higher education sector in England, the development field is facing reduced funding and is seeing a push for more accountability, results-based management and value-for-money. This has, invariably, resulted in an increased focus by research institutions and research funding agencies on the quality of research produced. Evaluation of research excellence and the debates around peer review, bibliometrics, and impact as a dimension of excellence are, therefore, growing in prominence in the research for development field.

However, designing research excellence evaluation processes is not simple. Excellence means different things to different people (Becker, Bryman, & Sempik, 2006; Martí & Villasante, 2009; Furlong & Oancea, 2005), suggesting a need for multi-layered and multi-dimensional evaluation approaches. In designing those approaches, those who evaluate research should keep in mind that research excellence evaluation systems should be transparent, comparable, fair, appropriate, less burdensome, and must be recognized by peers (Wooding & Grant, 2003, p. 15). Their design should also respond to a clearly stated evaluation purpose.

This review aimed at providing a glimpse of a fairly large and complex topic. There are, of course, many gaps and areas for further research, including:

- What is meant by research excellence in the research for international development field? While that question was the focus of much of this report, the literature found and reviewed indicates that much of the thinking on this issue has been incubated in institutional research settings in the global North.
- What is meant by each of the quality dimensions/criteria identified in this report? How can they be identified or measured in a research excellence evaluation process (i.e. what does relevance in interdisciplinary research mean? How can it be identified or measured?)?
- What are the advantages, disadvantages, and trade-offs of clearly delineating re-search "excellence" and research "quality"?

• How should the quality of interdisciplinary work and of revolutionary or discovery research be evaluated?
• How should research impacts - if considered a component of research excellence - be evaluated?

This literature review identified a few examples from the global North that can be further explored to inform how research excellence evaluation should be conducted in the development field. For example, while various components of England's REF have been contested, it is an example of an evaluation model that provides useful overarching features – research outputs, impact, and research environment - but does not go into the details of discipline-specific review panels, thus allowing flexibility in a process that is otherwise bound to criteria. It should be noted, however, that the views of institutions and researchers operating in the global South are largely missing in this report. Considering that research for international development is almost strictly about the global South, their opinions and experiences are essential. Future inquiries should delve further into Southern researcher and research institution perspectives on research excellence.

Finally, the questions raised in this paper have implications that go beyond academia. Evaluation of research excellence is often tied to funding and, through it, to the production of new knowledge and to innovation – both essential elements for societies, and not just academia, to evolve. As Boaz and Ashby note: *"After all, one of the strengths of research, compared to other sources of knowledge available to decision makers, should be that it is a quality assured product carried out to prepared standards. A broader notion of research quality should help researchers and research users to feel confident about the use of evidence in policy and practice"* (2003, p. 2).

Works Cited

Aagaard-Hansen, J., & Svedin, U. (2009). Quality Issues in Cross-disciplinary research: towards a two-pronged approach to evaluation. *Social Epistemology*, 165-176.

Aizenman, J., Edison, H., Leony, L., & Sun, Y. (2011). *Evaluating the Quality of IMF Research: A Citation Study*. Washington : Independent evaluation Office - IMF.

Andras, P. (2011). Research: metrics, quality, and management implications . *Reserach Evaluation,* 90-106.

Arnold, D. (2008). Cultural Heritage As a Vehicle for Basic Research in Computing Science: Pasteur's Quadrant and a Use-Inspired Basic Research Agenda. *COMPUTER GRAPHICS forum, Volume 27* (number 8), 2188–2196.

Arunachalam, S. (2009). *Social Science Research in South Asia An analysis of the published journal literature.*

Becker, S., Bryman, A., & Sempik, J. (2006). *Defining 'Quality' in Social Policy Research: Views, Perceptions and a Framework for Discussion*. Suffolk: Lavenham: Social Policy Association.

Boaz, A., & Ashby, D. (2003). *Fit for purpose? Assessing research quality for evidence based policy and practice*. Retrieved 2011, from ESRC UK Centre fro Evidence Based Policy and Practice: Working Paper 11: http://www.kcl.ac.uk/content/1/c6/03/46/04/wp11.pdf

Boix Mansilla, V., Feller, I., & Gardner, H. (2006). *Quality assessment in interdisciplinary research education.* Surrey: Research Evaluation.

Booth, D. (2011, April). *Working with the grain and swimming against the tide: Barriers to uptake os research finding on governance and public services in low-income Africa*. Retrieved July 7, 2011, from UK Department for International Development : http://www.dfid.gov.uk/r4d/PDF/Outputs/APPP/appp-working-paper-18.pdf

Bridges, D. (2009). Research quality assessment in education: impossible science, possible art?. *British Educational Research Journal*, 497-517.

Coryn, C. L. (2006). The Use and Abuse of Citations as Indicators of Research Quality. *Journal of MultiDisciplinary Evaluation*, 115-121.

Coryn, C. L., Hattie, J. A., Scriven, M., & Hartmann, D. J. (Dec 2007). Models and Mechanisms for Evaluating Government-funded Research: An International Comparison. *American Journal of Evaluation*, 437-457.

Coryn, C. (Pre-publication chapters from Coryn, C. L. S. (2013)). What is Social Science Research and why would we want to evaluate it? In C. Coryn, *Evaluating social science research:*

A handbook for researchers, instructors, and students. (pp. 1-32). New York, NY: Guilford.

David, M. E. (2008, 7, 1,). Research Quality Assessment and the Metrication of the Social Sciences. *European Political Science*, 52-63, Palgrave Macmillan Ltd.

Dictionary.com. ((n.d)). Retrieved November 24, 2011, from Dictionary.com Unabridged.: http://dictionary.reference.com/browse/research

Donaldson, S. (2009). In search for blueprint for an evidence-based global society. In S. I. Donaldson, A. Christina, & M. M. Mark, *What counts as credible evidence in applied research and evaluation practice?* (pp. 2-17). Thousand Oaks, California: Sage.

Donovan, C. (2007). The qualitative future of research evaluation. *Science and Public Policy*, 585–597.

Fielding, N. (2010). Elephants, gold standards and applied qualitative research. *Qualitative Research*, 123-127.

Furlong, J., & Oancea, A. (2005). *Assessing quality in Applied and Practice-based Educational Research*. Oxford: Oxford University Department of Educational Studies.

Graham, L. (2008). Rank and File: Assessing research quality in Australia. *Educational Philosophy and Theory*, 811-815.

Grant, J., Brutscher, P.-C., Kirk, S. E., Butler, L., & Wooding, S. (2010). *Capturing Research Impacts: A review of international practice*. Cambridge, UK: Rand Europe.

Groundwater-Smith, S., & Mockler, N. (2007). Ethics in practitioner research: an issue of quality. *Research Papers in Education*, 199-211.

Hammersley, M. (2008). Troubling criteria: A critical commentary on Furlong and Oancea's framework for assessing educational research. *British Educational Research Journal*, 747-762.

Higher Education Funding Council of England (2011). *Consultation on draft panel criteria and working methods*. England: HEFCE.

Higher Education Funding Council of England. (n.d.). *Higher Education Funding Council of England*. Retrieved January 25, 2012, from Research Excellence Framework : http://www.hefce.ac.uk/research/ref/

House, E. R. (2008). Blowback: Consequences of Evaluation for Evaluation. *American Journal of Evaluation*, Volume 29 (Number 4), 416-426.

Kenna, R., & Berche, B. (2011, June). Normalization of peer-evaluation measures of group research quality across academic disciplines. *Research Evaluation*, 107-116.

Lardone , M., & Roggero, M. (n.d.). *Study on monitoring and evaluation of the research impact in the public policy of Policy Research Institutes (PRIs) in the region*. Retrieved July 2011, from Evidence Based Policy Development Network: http://www.ebpdn.org/download/download.php?table=resources&id=3013

Laudel, G., & Glaser, J. (2006). Tensions between evaluations and communication practices. *Journal of Higher Education Policy and Management*, 289-295.

Liang, L., Wu, Y., & Li, J. (2001). Selection of databases, indicators and models for evaluating research performance of Chinese universities. *Research Evaluation*, 105-113.

London School of Economics. (2011). *Maximizing the impact of academic research*. Retrieved 2011, from Impact of Social Sciences: http://blogs.lse.ac.uk/impactofsocialsciences/

LSE Public Policy Group. (2011). *Maximizing the impacts of your research: A Handbook for Social Scientists*. London : LSE Public Policy Group.

Malterud, K. (2001). Qualitative research: standards, challenges, and guidelines. *The Lancet*, 483-488.

Marjanovic, S., Hanney, S., & Wooding, S. (2009). *A historical reflection on research evaluation studies, their recurrent themes and challenges*. Cambridge: RAND Corporation.

Martí, J., & Villasante, T. R. (2009). Quality in Action Researcj: Reflections for Second-Order Inquiry. *Systemic Practice and Action Research*, 383-396.

National Center for the Dissemination of Disability Research. (2005). *What are the standards for Quality Research - Focus technical brief*. Retrieved September 22, 2011, from National Center for the Dissemination of Disability Research: http://www.ncddr.org/kt/products/focus/focus9/

Nutley, S., Percy-Smith, J., & Solesbury, W. (2003). *Models of research impact: a cross-sector review of literature and practice*. London: Learning and Skills Research Centre.

O'Gorman, L. (2008, January). *The (frustrating) state of peer review*. Retrieved July 2011, from International Association for Patter Recognition Newsletter: http://iapr.org/docs/newsletter-2008-01.pdf

O'Neil, M. (2002). Commentary: We May Need a New Definition for Research Excellence. *University Affairs*.

Organisation for Economic Co-operation and Development. (1997). *The evaluation of scientific research: Selected Experiences*. Paris: OECD.

Petit-Zeman, S. (2003, January 16). Trial by peers comes up short. *The Guardian*, http://www.guardian.co.uk/science/2003/jan/16/science.research

RAND Corporation. (2011, November). *RAND's Standards for High-Quality Research and Analysis*. Retrieved January 2012, from RAND Corporation: http://www.rand.org/standards/standards_high.html

REF 2014. (July 2011). *REF 2014: Assessment Framework and guidance on submission*. UK: HEFCE.

Rodrik, D. (2011, June 10). *Dani Rodrik's weblog: Unconventional thoughts on economic development and globalization*. Retrieved June 2011, from A rejection letter I would like to receive from a journal one day: http://rodrik.typepad.com/dani_rodriks_weblog/2011/06/a-rejection-letter-i-would-like-to-receive-from-a-journal-one-day.html

Roebber, P. J., & Schultz, D. M. (2011, April 12). *Peer Review, Program Officers and Science Funding*. Retrieved July 25, 2011, from PLoS One : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3075261/

Rons, R., De Bruyn, A., & Cornelis, J. (2008). Research evaluation per discipline: a peer-review method and its outcomes. *Research Evaluation*, 45-57.

Rowland, F. (2002, October). *The Peer Review Process*. Retrieved July 2011, from Joint Information Systems Committee : http://www.jisc.ac.uk/uploaded_documents/rowland.pdf

Saari, E., & Kallio, K. (2011). Developmental Impact Evaluation for Facilitating LEarning in Innovation Networks. *American Journal of Evaluation*, 227-245.

Sastry, T., & Bekhradnia, B. (2006). *Using metrics to allocate research funds: initial response to the Government's consultation proposals*. Oxford: Oxford Higher Education Policy Institute (HEPI).

Seymour, F. (2011, June 27). *Does the pressure for impact compromise research?* Retrieved June 2011, from Forests Blog: http://blog.cifor.org/3439/does-the-pressure-for-impact-compromise-research/

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in Qualitative Evaluation: A framework for assessing research evidence*. UK Government Chied Social Researcher's Office.

Stephen, C., & Daibes, I. (2010). *Defining features of the practice of global health research: an examination of 14 global health research teams.* Retrieved July 4, 2011, from Global Health Action : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2903310/

Tatavarti, R., Sridevi, N., & Kothari, D. (2010). Assessing the quality of university research – the RT factor. *General Science* , 1015-1019.

Tijssen, R. J., Mouton, J., van Leeuwen, T. N., & Boshoff, N. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation , Volume 15* (number 3), 163-174.

Tijssen, R. (2003). Scoreboards of research excellence. *Research Evaluation*, 91-103.

Wagner, C. S., Roessner, J., Bobb, K., Thompson Klein, J., Boyack, K. W., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A re-

view of the literature. *Journal of Infometrics* , 14-26.

Ware, M. (2011). Peer review: recent experience and future direction. *New Review of Information Networking*, 23-53.

Webb, C. (1993). Feminist research: definitions, methodology, methods and evaluation. *Journal of Advanced Nursing* , 416-423.

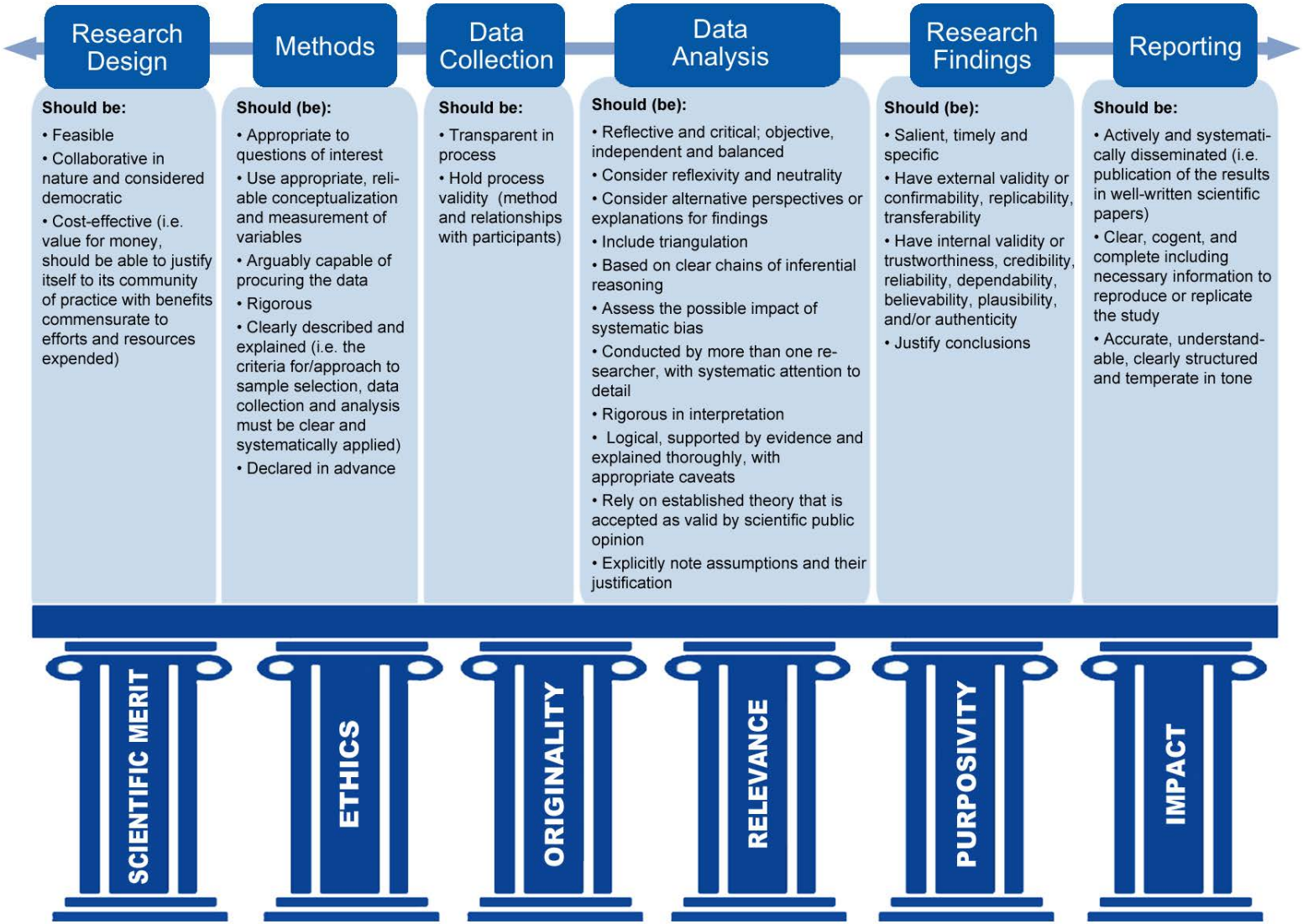
Wooding, S., & Grant, J. (2003). *Assessing Research: The Researcher's View.* Retrieved from RA Review: http://www.ra-review.ac.uk/reports/assess/AssessResearchReport.pdf

Wu, H., Ismail, S., Guthrie, S., & Wooding, S. (2011). *Alternatives to Peer Review in Research Project Funding.* Cambridge, UK: RAND Corporation .

Yates, L. (2005). Is Impact a measure of Quality? Some Reflections on the Research Quality and Impact Assessment Agendas. *European Educational Research Journal , Volume 4* (Number 4), 391-403.

Yule, M. (2010). *Assessing Research Quality.* Ottawa: International Development Research Centre - Peace Conflict and Development.

Annex 1: Common Conceptual Elements and Criteria of Research Excellence in the Research Process



List of acronyms

ERiC - Dutch Evaluating Research in Context

HEFCE – Higher Education Funding Council of England

IDRC – International Development Research Centre

IMF – International Monetary Fund

OECD – Organization for Economic Cooperation and Development

PART - Program Assessment Rating Tool

RAE – Research Assessment Exercise

RAISS - Arthritis Research Campaign Impact scoring system

REF – Research Excellence Framework

RQF - Research Quality and Accessibility Framework