Arts and humanities research evaluation: No metrics please, just data[1]

Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK
Maria M. Delgado, School of English and Drama, Queen Mary University of London, UK

[*Journal of Documentation*, vol. 71, no. 4 (July 2015), 817-833]; http://dx.doi.org/10.1108/JD-02-2015-0028

Purpose: To make an explicit case for the use of data with contextual information as evidence in arts and humanities research evaluations rather than systematic metrics.
Design/methodology/approach: A survey of the strengths and limitations of citation-based indicators is combined with evidence about existing uses of wider impact data in the arts and humanities, with particular reference to the 2014 UK Research Excellence Framework.
Findings: Data are already used as impact evidence in the arts and humanities but this practice should become more widespread.
Practical implications: Arts and humanities researchers should be encouraged to think creatively about the kinds of data that they may be able to generate in support of the value of their research and should not rely upon standardised metrics.
Originality/value: This paper combines practices emerging in the arts and humanities with research evaluation from a scientometric perspective to generate new recommendations.

Introduction

The use of metrics derived from academic citations in formal and informal research evaluations has a long and controversial history. The birth of the Science Citation Index in 1964 made it practical to use citation-based metrics in research evaluation for the first time and various metrics have since then been used to aid the evaluation of journals, articles and authors (Garfield, 1979). The fundamental principal underlying citation analysis is that scientists tend to acknowledge prior research informing their work by citing it and that this is normative behaviour in science (Merton, 1973). This principal has been widely criticised on the basis that citations can be negative and influenced by factors other than the quality or contribution of the cited article (Brooks, 1986; MacRoberts and MacRoberts, 1996; Seglen, 1998) and so citation counts are not direct and unbiased measures of the contributions of articles to future research. Academics may also be concerned that metrics-based evaluations use numbers that they do not recognise as reflecting themselves (Day, 2014). Nevertheless, evidence that indicators based on citation counts in some subject areas tend to positively correlate with citation counts (e.g., Franceschet and Costantini, 2011) has been used to support their use in research evaluation. This use would typically be to inform expert judgements about research quality rather than to replace them (Moed, 2006; Nederhof and Van Raan, 1993; Warner, 2000). For example, if the outputs of one or more researchers need to be rated then the

citation indicators could be used as a starting point for the evaluation, as an alternative data source to cross-check against the initial human ratings, or as an additional source of evidence for marginal cases.

A rationale for the use of citation counts in research evaluation despite some citations being negative and despite the presence of some biasing factors is that, over a reasonably large collection of publications, the biases may tend to even out (van Raan, 1998). This would explain the positive correlations found between expert judgements and citation metrics (see Appendix). Nevertheless, since academics alter their patterns of research in response to the criteria set for assessment, at least in the UK (Moed, 2008), and bibliometric indicators may favour men (HEFCE, 2011), the implications of any changes on behaviour need to be considered.

Metrics have already been used to support peer judgements in evaluations in several countries. In the UK Research Excellence Framework (REF) 2014 all 36 sub-panels (i.e., subject groupings) were offered the use of externally gathered Scopus citation counts together with annual field averages. This option was taken by: all sub-panels of the health and life sciences panel (Clinical Medicine; Public Health, Health Services and Primary Care; Allied Health Professions, Dentistry, Nursing and Pharmacy; Psychology, Psychiatry and Neuroscience; Biological Sciences; Agriculture, Veterinary, and Food Science), and some of the natural and formal sciences and engineering panel (Earth Systems and Environmental Sciences; Chemistry; Physics; Computer Science and Informatics) but not mathematics nor the three engineering sub-panels. In engineering, there may have been a concern that citation counts would favour pure research over applied research, whereas the latter is highly valued in the discipline. In mathematics, citations may be seen as completely irrelevant to the quality of mathematics. For example, the highest prize in mathematics, the Fields Medal, was given to a mathematician, Maryam Mirzakhani, in 2014 who had received relatively few citations (e.g., 536 in Google Scholar by February 2015). In the social sciences panel, only the Economics and Econometrics sub-panel opted to be given the citation metrics and none of the arts and humanities sub-panels chose them (HEFCE, 2014). Citation counts were not central to the judgments in the sub-panels that used them, however, but were mainly used to help resolve disagreements between evaluators, at least in Main Panel A (REF2014, 2015a), and Journal Impact Factors were completely ignored (REF2014, 2015a).

Excellence in Research Australia (ERA) has taken a very similar approach, providing journal article citation data to inform peer judgements of quality for Mathematical (except pure), Physical, Chemical, Earth, Environmental, Biological, Medical and Health, Agricultural and Veterinary Sciences but not for Information and Computing Sciences (ERA, 2015). Citation data is also provided for Engineering and Technology (except computing), but not for Built Environment and Design. Within the social sciences, arts and humanities, citation data is only provided for Psychological and Cognitive Sciences. In contrast, the New Zealand Performance-Based Research Fund (PBRF) assesses individuals on the basis of a portfolio of work submitted, without disciplinary restrictions. All researchers are explicitly permitted to submit evidence of the positive citations (and positive reviews) that they have received as long as they interpret them and justify them as being positive (PBRF, 2013).

Despite the New Zealand example, researchers may be forgiven for thinking that if they reject citation counts then the main source of evidence of the value of their work is a list of outputs and perhaps also evidence of peer recognition, such as editorships, the delivery of keynotes, consultancies and awards. This is limiting, however, since evidence about the impact of work can help to make a stronger case

for its value both within and beyond the academic sector. In addition, for a number of disciplines where practice as research is a recognised mode of investigation, the relationship between the dissemination of findings and impact cannot be easily unpicked. In the creative and performing arts, it is not unusual for impact to feature as part of the whole research journey rather than as an annex conceived after the findings have been disseminated. Where performances, installations, and exhibitions are placed in front of a public or have involved wider users, feedback from those stakeholders on the process and/or outcome as it relates to an explicit or stated research intention can productively form part of the process of documenting the project's significance, originality and rigour. Impact is thus often embodied within the research rather than generated subsequently (see REF2014, 2015d, Sub-panel 35, para. 21). This article reviews evidence about disciplinary differences in the use of citations for research evaluation and gives examples of wider uses of data in arts and humanities evaluations. It finishes with recommendations for the wider use of data as evidence for the impact of arts and humanities research and argues for a careful use of terminology that recognizes the place of non-systematic data rather than systematic citation (or other: Cronin, 2014; Priem, Taraborelli, Groth, and Neylon, 2010) metrics in the assessment of arts and humanities research.

Impact vs. engagement: The importance of context

The value that arts and humanities research gives to society is not always as transparent or visible as for other types of research, such as medical or technological innovations. As a result, there is less of a public consensus about the benefits of arts and humanities research in general, and a more concerted case has to be made for justifying the public money that it receives to support research. In particular, it is impossible to demonstrate the socio-economic impact of some arts and humanities scholarship because it is valuable in other ways, some of which are impossible to measure or even estimate in any meaningful way (Belfiore and Upchurch, 2013). The increasing use of evidence-based policy therefore risks marginalising the highly subjective experience of engaging with the arts (Belfiore and Bennett, 2008, 5-9).

The humanities may have value in five primary ways: for their insights into meaning-making and knowledge; their distance from practical applications; their contribution to happiness; their contribution to democracy; and for their own sake (Small, 2013). For example, the value of the humanities within medical education has been justified on the basis that it can help clinicians to cope with varied and complex forms of evidence and knowledge from which a decision must be made (Belling, 2010). In addition, each individual area of the arts and humanities probably has its own distinctive type of contribution and justification for the value of its work (e.g., Bate, 2011). For instance, religious history could inform understandings of the recent rise of Christian and Islamic fundamentalism and this may help to generate better informed policy today (Wolffe, 2011). There is perhaps more emphasis on the value of arts and humanities in education rather than the outputs of researchers in comparison to other disciplines. One argument even makes the case that the humanities aid democracy by educating the nation to engage more effectively in the political process (Nussbaum, 2012). Another recognises the influence of linguistics on the widespread changes introduced to the education system in England in the 1980s and 1990s (Hudson, 2007). Furthermore, the rich tradition of research in applied arts practice, encompassing such areas as music therapy, community dance, and social theatre, has produced significant findings about the social, medical and emotional benefits of such practices of engagement to civic society, effectively

extending the cultural sector's influence into the fields of welfare and social justice (see, for example, Heritage, 2005; Oldfield, 2006; Thompson, 2009).

Although citations seem to be rarely used within the arts and humanities, data is routinely used to demonstrate value in a number of contexts. The UK Arts and Humanities Research Council (AHRC), for example, gives guidelines to the owners of its funded projects about how to self-evaluate their progress (AHRC, 2015). They differentiate between three separate things that directly or indirectly occur as a result of the projects:

- *Outputs*: Tangible things produced.
- *Outcomes:* Changes in "behaviour, skills, status and level of functioning" of participants.
- *Impact:* Fundamental changes in "organisations, communities or systems".

Although the outputs are trivial to identify and impacts are likely to be hard or impossible to measure, information about outcomes can be collected through interviews, questionnaires, focus groups and other social research methods, generating a mix of qualitative and quantitative data. This seems to be recognised practice in areas, such as music therapy (e.g., Heaney, 1992) and the use of the arts in pedagogy (Kontos and Naglie, 2007) or rehabilitation (Gussak, 2006; Johnson, 2008; Vacca, 2004), where there is a clear and measureable goal. In the UK REF, however, all submitted groups of researchers were expected to submit self-contained impact case studies that described how their research translated into non-academic impacts, as explained in more detail in the next section. It is therefore useful to distinguish between outcomes and impacts within and outside of the scholarly community.

- *Scholarly outcomes:* Outcomes reflected in other scholars, inside or outside of the discipline.
- *Wider outcomes:* Outcomes reflected outside of the scholarly community.
- *Scholarly impact:* Impact on scholarship inside or outside of the discipline.
- *Wider impact:* Impact outside of the scholarly community.

An example of a type of data commonly used in the arts in contexts where there is not a clearly measurable external goal, audience sizes are routinely used to evaluate the reach of entertainment outputs. Numbers alone are insufficient, however, and must be interpreted in context to be translated into evidence about the level of engagement or transformation within the audience (Holden, 2004). Arts Council England (ACE), for example, prioritised "developing arts opportunities for people and places with the least engagement" in its 2011-15 plan (Arts Council England, 2013). Reaching a smaller audience in the 71 local authorities highlighted by ACE as having the lowest arts engagement might therefore count as more significant in terms of wider impact than a larger audience in a metropolitan centre or area judged as having a more pronounced level of cultural engagement. The Appreciation Index or AI is used by organisations such as the BBC to register the audience enjoyment levels of radio and television programmes (BBC, 2014). Again here the focus is on the quality of experience rather than the quantity of viewers or listeners. Hence audience sizes could be evidence of wider outcomes in the above terminology but contextual information would be needed to turn it into evidence of wider impact.

Within the UK REF2014, each submission had to include two or more case studies that described how the research of the group had had an impact. The impact guidelines state that, "The onus is on submitting units to provide appropriate evidence within each case study of the particular impact claimed. The REF panels will provide guidance, in the panel criteria documents, about the kinds of evidence

and indicators of impact they would consider appropriate to research in their [areas], but this guidance will not be exhaustive" (REF2014a, 2012, para. 164). In the arts and humanities Main Panel D the case study evaluation guidance was that, "public engagement may be an important feature of many case studies, typically as the mechanism by which the impact claimed has been achieved" (REF2014, 2012b, para. 83) and the likely range of impacts for arts and humanities research was listed as civil society, cultural life, economic prosperity, education, policy making, public discourse and public services (REF2014, 2012b, Table D1). Examples of a range of qualitative and quantitative sources of evidence that could be provided to support an impact claim were given (REF2014, 2012b, Table D2). The "indicators" listed included publication and sales figures, external funding, evidence of use of educational materials, tourism data, and business growth figures, such as income, or employment. The other examples of impact evidence included critiques or citations from users, public engagement data (including numbers and descriptions), policy engagements, independent testimony and formal evaluations.

The range of data provided as part of the impact component of REF2014 testifies to an understanding of the ways in which numbers can be used to demonstrate the reach and significance of the impact deriving from research. A consideration of the impact case studies presented to the Music, Drama, Dance and Performing Arts sub-panel 35 (see http://results.ref.ac.uk/Results/ByUoa/35), demonstrates the widespread use of data to underpin and illustrate the claims being made. This includes: the amount raised in private donations used to fund a music therapy centre; audience numbers for broadcasts, creative works and performances; user numbers for software; participant numbers for community projects; visitor numbers for museums and installations; sales and/or download figures for CDs and DVDs, books, scholarly editions, film admissions, magazine articles or print music; membership numbers for artistic initiatives; quantifiable visitor comments and hosted residencies; website hits, tweets and social media presence; quantity of students reached beyond the host HEI.

This is not to suggest that quantity is necessary or sufficient to demonstrate impact. Quantity may suggest reach but does not testify to significance. The latter might be evidenced by a major change in cultural strategic thinking or policy emerging from a single consultancy with a high-ranking politician or civil servant. This is how the contextualising narrative of each case study functions to position, frame and explain the data. The sub-panel 35 report signals that the strongest impact case studies demonstrated 'a clear awareness of users, audiences and beneficiaries' with data rather than generalized statements used to map the impact (REF2014, 2015d, Sub-panel 35, para. 57-8). When appropriately contextualised, 'meaningful (in terms of quality and quantity) data from participants and beneficiaries on impact deriving from research' made a difference (REF2014, 2015d, Sub-panel 35, para. 58).

Environment data: A case study from Music, Drama, Dance and Performing Arts

The observations here relate to the fact that while there is unease about the use of metrics as a mode of 'measuring' the excellence of research produced in the UK's Higher Education Institutions (HEIs), the rich array of data presented as part of REF2014 demonstrates that the arts and humanities sector are comfortable with deploying numbers (albeit framed as data rather than metrics) to present a case

about the excellence of their research cultures. When considering the Environment component of the process, making up 15% of the overall score of a unit's assessment, data was similarly submitted across a range of areas and used to inform rather than determine judgements. Units of Assessment (UoAs, i.e., individual submissions from universities) were asked to include data on the total number of doctoral degrees awarded and research income from the external sources listed on para. 171 of Assessment framework and guidance on submissions document (REF2014, 2012a, p. 31), as verified with the Higher Education Statistics Agency (HESA). This HESA data includes all activity associated with a submission, including activity associated with staff that were not selected for inclusion. Thus it could include data associated with staff who had since left the unit, staff who were judged by HEIs not to have enough high quality outputs to be included in their university's submission, staff who were not eligible, or even staff from other areas that were transferred for strategic reasons. The HESA data would be most useful if mapped to the research achievements delineated through the outputs submitted to the REF and the projects and collaborations that demonstrate the vitality and sustainability of the submitting department or unit. As with the impact case studies, strategic judgements were apparently made by HEIs to utilise selective data that underpinned the narratives they were presenting about their research achievements over the period of assessment.

The number of doctoral degrees awarded, for example, has to be read within a context: what information has been provided about robust institutional support through clearly-articulated 'procedures and finance for field-work, travel and conferences; training programmes that reflected the research imperatives of the discipline, rather than purely generic needs; and strong external links', for example in relation to engagement with the creative industries and/or practice as research (REF2014, 2015d, Sub-panel 35, para. 67). This data might provide some indicators of vitality and sustainability but quantity alone is not enough to demonstrate an excellent research environment. Evidence of completion rates (not always provided), 'awards from funding bodies, performance opportunities, prizes, publications and appointments, PGR students taking a leading role in organising seminars, conferences, and workshops, and in promoting and disseminating innovative research through electronic journals' and other publication initiatives provide a much wider picture of the richness of the unit's research culture (REF2014, 2015d, Sub-panel 35, para. 67). The number of doctoral degrees awarded might not reveal whether postgraduate students are fully integrated into the wider research culture of the HEI in a sustainable way; or how smaller or emerging units provide a vital structure for postgraduate supervision and training. In addition, at a time when considerably less than 50% of graduate students in this the arts and humanities are securing permanent or fixed-term employment in Higher Education (Rothman 2014; Renfrew and Green, 2014), a demonstration of the employment destinations of graduate students would be useful. What are our graduate students going on to achieve and how are we preparing them for the challenges of using their knowledge and training both within and beyond academia?

With regard to external research income, the REF uses HESA definitions of research income that come from a particular set of funding sources. Research in the creative arts, especially practice as research, is often funded from bodies that are not captured by HESA data. The importance of this income (from national Arts Councils, national and local government bodies and think-tanks, orchestras and opera houses, recording companies, broadcasters, distributors, exhibitors, promoters

etc.) (see REF2014 2015d Sub-panel 35 para. 71) was often clearly related to outstanding research projects in the Unit's Environment template. The standing of non-HESA data is something that merits careful consideration in the cluster of creative arts disciplines (art, design, music, drama, dance and performing arts).

Furthermore, a percentage of the research in these disciplines prioritises participatory and collaborative practices with a number of practitioners contributing to the underpinning research. In some cases it is hard to disaggregate which income belongs to which practitioner. This needs to be factored into any amendments to the metrics required for future assessment exercises. Data may show how successful a unit has been in obtaining income but not how productive has this been. For example, whilst some types of research need funding to happen (e.g., large scale performances; musical innovations requiring new equipment), for others (e.g., theoretical research or creative writing) may not. The same is probably true in the sciences, except with larger amounts of money. Data alone didn't indicate how research funding had strengthened the staff base, brought in early career researchers, generated new areas of research or fed into particular submitted outputs. Hence there are dangers with using income as a definitive metric. Data needs to be read within an informed context. While there was a varied use of data in REF2014 – and, moving forward, a greater standardization and guidelines on capturing data would be useful – it is imperative to ensure that whatever systems are devised for REF2014's successor, which likely to be in 2020, they do not prove restrictive and counterproductive. The real benefits of research assessment exercises lies in their initiation of a process which allows departments and disciplinary clusters to reflect on what they do and how they do it. All four Main Panel reports located peer-review founded on expert judgement supported by appropriate quantitative data as 'the heart of the assessment process' for REF2014 (REF2014, 2015a-d, quote taken from 2015d, Main Panel D, para. 9).

Citations to arts and humanities research

An important problem for any citation-based evaluations of arts and humanities research is that artists produce a wide variety of types of outputs, including musical scores and instrumental performances, software design and dance performances, which are not naturally citable. The report from sub-panel 35 delineates over 34 different types of output submitted for REF2014 assessment: 'advisory reports and evaluations, books (authored and edited), chapters in books, journal articles, published conference papers, electronic resources and publications, exhibition catalogues, translations and scholarly editions, compositions and musical scores, creative writing (libretti, film scripts, radio plays, novels, short stories, stage plays), databases, grammars, patents, digital and broadcast media, performances, films, video and media presentations, installations, designs and exhibitions, software design and development, working papers' (REF2014 2015d, Sub-panel 35, para. 38). Only 29.5% of the research outputs submitted to UoA35 came in article form (compared to 37.7% across Main Panel D, 99.5% across Main Panel A, 94.4% across Main Panel B and 80.9% across Main Panel C (REF2014 2015a-d).

About 42% of the outputs submitted to sub-panel 35 in the area of music, for example, were in non-text media (REF2014 2015d, Sub-panel 35, para. 11); the figure for the Drama, Dance and Performing Arts cluster (that includes film and screen media) is lower at 22% (REF2014 2015d, Sub-panel 35, para. 22). This is not to suggest that practice as research does not draw on research practices that combine more established methodologies from the humanities and social sciences

(Brown, 2002). Indeed, critical or theoretical perspectives from phenomenology, cultural materialism, human geography and sociology often underpin the research imperatives of critical practice as both a methodology and a mode of dissemination (Stige, 2005; Nelson, 2013; Delgado and Bottoms, 2010). In addition, monographs are particularly important in the humanities – 26.7% of the outputs submitted to Main Panel D were some form of book (authored, edited or a scholarly edition) and whilst these can be easily cited, citations from books are difficult to find because current citation indexes are dominated by academic journals and the book-based coverage of the Web of Science (Torres-Salinas, Robinson-Garcia, Campanario, and López-Cózar, 2014) and Scopus (Kousha and Thelwall, in press) are not comprehensive enough for evaluation purposes. Moreover, although some publishers peer review books and select authors carefully, they may still favour more popular research topics in order to guarantee sales. There are also problems with internationality of coverage that affect the humanities more than the sciences (Torres-Salinas, Robinson-Garcia, Campanario, and López-Cózar, 2014). Less than 1% of outputs submitted to Main Panel D were conference contributions, however (REF2014, 2015d, Sub-panel 35, Table 4). These seem to have little role in arts and humanities research cultures as a primary mode of disseminating findings and often function in a formative capacity to share work in progress.

Perhaps a more fundamental problem is that whilst scientists and social scientists can claim to some extent to be building a hierarchical knowledge structure in which it is important to cite previous work to demonstrate the position of the new work (Merton, 1973), this is not true for the arts and humanities. Instead, creativity is valued in the arts, which is to some extent the opposite of hierarchical knowledge construction, and humanities scholars may cite to demonstrate the originality of their work rather than their contribution to the body of knowledge (Hellqvist, 2010). In addition, they may cite from relatively unrelated fields in an attempt to broaden their potential audience (Hyland, 2004) and, for this, what they cite may be less important than who they cite. Some fields, such as cultural history, also need to cite ancient primary sources extensively and this may tend to suppress citations to contemporary research, making citation counts less useful for contemporary research evaluation in the field.

High citation counts may be less desirable in the arts and humanities than elsewhere because of the nature of highly cited areas. In science, hot topics may be the most cited and contributions to these hot topics, sometimes known as the research front (Åström, 2007; Boyack and Klavans, 2010), may be highly valued so that their high citations reflect peer values. In contrast, in the humanities, research in fashionable areas may be treated with some suspicion and their high citation counts therefore reflect the opposite of the wider community's opinion. This could be because humanities scholars are more specialised and less prone to contributing to teams than are scientists (Larivière, Gingras, and Archambault, 2006). Moreover the core humanities output, the monograph, seems not to benefit from collaboration (Thelwall and Sud, 2014), whereas the core science and social science outputs, journal articles do (Didegah and Thelwall, 2013; Glänzel, 2002). Thus, arts and humanities scholars probably have relatively few opportunities and incentives to collaborate on a different research topic. Nevertheless, one study with Austrian humanities scholars did not find clear evidence that they valued collaborative research differently from solo research (Ochsner, Hug, and Daniel, 2014). Nevertheless, it is difficult to see, for example, how a scholar of Old Norse could contribute to a more fashionable related area, such as Mandarin, without essentially

starting almost from the beginning. In the humanities, controversial or poor scholarship may also be repeatedly challenged and hence accrue high citation counts. Such scholarship may even have ongoing value as a convenient source against which to make the case for a generally accepted argument. Whilst poor or incorrect research can also be cited to be challenged in science, and can be highly cited (Ioannidis, 2005), this may be less common because, at least in principle, a scientific fact only needs to be discredited once, and facts are probably more easily stated rather than argued for.

Finally, a recent initiative to ask arts and humanities scholars from various disciplines which indicators could be used to help assess the quality of research has shown that there can be a field-specific consensus about this (Ochsner, Hug, and Daniel, 2014). Based on a survey of Swiss researchers in the fields of German literature studies, English literature studies, and art history, each area reached agreement about at least one relevant indicator. Nevertheless, none of the indicators were chosen by all three disciplines and the number of indicators chosen varied from one (English literature studies chose publications as an indicator of international exchange) to seventeen (art history, which chose, amongst other indicators, the number of sources, materials and original works used in publications or presentations as an indicator of rich experience with sources). Moreover, the very many groups of types of indicators suggested in this study is further evidence that citation counts alone are insufficient for humanities research evaluations, particularly given that the participants did not select citation counts and agreed that the selected indicators alone were insufficient for evaluations.

Overall, then, any set of citations to a typical collection of arts and humanities outputs are likely to be much less comprehensive than a corresponding set of citations to natural, life, formal or social science outputs and any citation counts derived from them would lack a straightforward connection to disciplinary research goals. In the terminology above, citations reflect scholarly outcomes and so citation counts could be scholarly outcome indicators. In contrast to other areas of scholarship, citation counts reflect scholarly outcomes more partially because of the technical and theoretical citation counting limitations discussed above. In science, strong correlations between citation counts and peer review quality scores (see appendix) in conjunction with the hierarchical nature of science and to some extent the social sciences, suggest that appropriately normalised citation counts could also be reasonable indicators of scholarly impacts. Both strands of this argument are weaker for the arts and humanities, however, because correlations with other scholars' quality judgements are weaker and there is not a strong theoretical link between citing and importance to scholarship. Hence citation counts are likely to be, at best, a weak indicator of scholarly impact in the arts and humanities.

Interestingly, the AHRC's 2015-2016 Delivery Plan observes that: 'The BIS/Elsevier study of the *International Comparative Performance of the UK Research Base* (2011) calculated that the field-weighted citation impact (a key measure of quality adjusting for the different fields of research in different countries) for UK humanities increased from 1.0 (the world average) for citations 1996-2000 to 1.25 for 2006-10. This measure of quality for UK humanities research is now higher than for the US, which scores 1.1. The arts and humanities therefore play a full part in the UK's reputation for exceptional distinction and value-for-money in research' (AHRC 2014 para. 3.1). Nevertheless, although the UK arts and humanities seem to attract relatively many citations, none of the arts and humanities sub-panels opted to use citation data to inform their judgements, even in the marginal way that they were

used by Main Panel A, and so it seems likely that citation data would rarely be valued more highly outside the UK for arts and humanities research.

Conclusions: The case for non-systematic data not metrics

Although qualitative or quantitative data is recommended for evaluations of funded AHRC projects in the UK, and there are presumably similar requirements elsewhere, no data is systematically collected in the REF to support evaluations of the researchers' outputs throughout the arts and humanities because none of the arts and humanities sub-panels chose to use citation counts. As argued above, there is not currently enough evidence to be reasonably sure that citation count data would help to improve the accuracy of peer review judgments about arts and humanities research because it would be much weaker as an indicator of scholarly impact. There are also concerns that the introduction of citation data would serve as a perverse incentive, handicapping less fashionable areas of research and so the overall effect of systematically introducing citations could be negative. This is not to posit an argument that pushes for the exceptionalism of the arts and humanities, but is a pragmatic recognition that no evaluation method should be responsible for changing the behaviour of researchers to the detriment of the range and reach of their research.

Nevertheless, since data can help in some arts and humanities evaluations of funded research projects, it seems logical to encourage this approach in other types of evaluations, such as REF impact case studies, careers (e.g., for appointments, promotions and tenure) and departmental evaluations (and similar). In these contexts, the diversity of the arts and humanities and the need for context to interpret results suggest that it is unlikely that it would be possible to generate a set of types of data that all should report, or even that data from two sources would be comparable. For example, although appropriately normalised citation counts may be broadly comparable even between different disciplines, it would be wholly inadequate to compare even audience numbers between two different performances of the same play because of the need to translate the raw outcome data (audience sizes) about the performances into at least more detailed outcome information about behaviour changes and, ideally, information about the likely overall impact. Thus a drive towards standardisation (as has occurred, with some justification, for citation data) would be counterproductive. Instead, those evaluated should be free to choose their own data to report but should accompany the data with a narrative to explain the context and significance of the numbers in their case. The evaluation of that data and narrative would therefore inevitably be subjective and made by human judges rather than an algorithm. This is in agreement with a previous claim that systematic methods to evaluate arts impact, such as the toolkits of standard methods sometimes found in the social sciences, are not suitable because they would necessarily oversimplify the effects of artistic engagement (Belfiore and Bennett, 2010).

The advantage of the more widespread use of contextualised data in arts and humanities research evaluations would be twofold: evaluations would arguably be more accurate because they would be not just based upon narratives but would also contain supporting evidence; and that the researchers themselves would be driven towards thinking more precisely about the types of outcomes and impacts that their work produces, which seems likely to help them focus on the most impactful type of research within their scope. The disadvantages, however, would be the perverse incentive to do types of research for which data was more readily available and the

time taken to gather such data. This might be a substantial amount of time if, for example, questionnaires needed to be designed, distributed and evaluated. A final drawback is that some areas of the humanities may have no outcomes but primarily engage in discipline building without significantly engaging with a wider audience. These would be disadvantaged in evaluations in which the use of data is encouraged and so would have to compensate by developing arguments to demonstrate their value to society as well as arguments about why this value cannot be reflected in any kind of data.

While these recommendations relate directly the findings above about arts and humanities research, they may also be applicable to some other subject areas that consider traditional academic metrics to be unhelpful.

As a final point, the language used in the debate about the need for data is important. Arts and humanities researchers may be justifiably wary of any attempt to make them use metrics because of the standardisation connotations of the term with both citation counting and the drive to translate social and cultural benefits into economic terms. A particular problem with the term metric, and also to a lesser extent with the weaker term indicator, is that without tedious repetition of the *what* that is measured, it is easy for casual users to mistake metrics or indicators as being measures of research quality rather than measures of something else that is assumed to relate in some way to research quality. Thus, it is common to argue that metrics don't work with examples of why they do not measure research quality. In contrast, the term data (or evidence) does not carry the same connotations and a call to provide data in support of claims of research outcomes or impact is intuitively more reasonable and therefore more likely to succeed.

Acknowledgements

Thank you to Elizabeth Westlake for discussions, advice and calculations relating to REF sub-panel issues. Thank you also to members of sub-panel 35 for many relevant discussions during the deliberations.

References

AHRC (2014), "Arts and Humanities Research Council Delivery Plan 2015-2016", http://www.ahrc.ac.uk/News-and-Events/Publications/Documents/AHRC%20Delivery%20Plan%202015-16%20%28A%29.pdf (accessed 13 February 2015)

AHRC (2015), "Understanding your project: A guide to self-evaluation", http://www.ahrc.ac.uk/What-We-Do/Build-the-evidence-base/Pages/Self-evaluation.aspx (accessed 13 February 2015)

Aksnes, D. W., and Taxt, R. E. (2004), "Peer reviews and bibliometric indicators: a comparative study at a Norwegian university", *Research Evaluation*, Vol. 13 No. 1, pp. 33-41.

Anderson, R. C., Narin, F., and McAllister, P. (1978). Publication ratings versus peer ratings of universities. Journal of the American Society for Information Science, 29(2), 91-103.

Arts Council England (2013), "The Arts Council Plan 2011-15", http://www.artscouncil.org.uk/media/uploads/pdf/Arts_Council_Plan_2011-15.pdf (accessed 13 February 2015)

Åström, F. (2007), "Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 7, pp. 947-957.

Bate, J. (Ed.). (2011), *The public value of the humanities*, Bloomsbury Academic, London.

BBC (2014), "Annual report 2013/14", http://www.bbc.co.uk/annualreport/2014/glossary (accessed 13 February 2015)

Belfiore, E., and Upchurch, A. (Eds.). (2013), *Humanities in the twenty-first century: Beyond utility and markets*, Palgrave Macmillan, Basingstoke, UK.

Belfiore E. and Bennett, O. (2008), *The Social Impact of the Arts: An Intellectual History*, Palgrave Macmillan, Houndmills.

Belfiore, E., and Bennett, O. (2010), "Beyond the “Toolkit Approach”: arts impact evaluation research and the realities of cultural policy-making", *Journal for Cultural Research*, Vol. 14 No. 2, pp. 121-142.

Belling, C. (2010), "Commentary: Sharper instruments: On defending the humanities in undergraduate medical education", *Academic Medicine*, Vol. 85 No. 6, pp. 938-940.

Boyack, K. W., and Klavans, R. (2010), "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 12, pp. 2389-2404.

Brooks, T. A. (1986), "Evidence of complex citer motivations", *Journal of the American Society for Information Science*, Vol. 37 No. 1, pp. 34-36.

Brown, C. D. (2002), "Straddling the humanities and social sciences: The research process of music scholars", *Library & Information Science Research*, Vol. 24 No. 1, pp. 73-94.

Cronin, B. (2014), "Scholars and scripts, spoors and scores" in Cronin, B. and Sugimoto, C.R. (Ed.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*:, MIT Press, Cambridge, MA, pp. 3-21.

Dancy, C. P., and Reidy, J. (2004), *Statistics without maths for psychology*, Pearson Education Limited, Harlow.

Day, R. E. (2014), "The data—it is me!" in Cronin, B. and Sugimoto, C.R. (Ed.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*:, MIT Press, Cambridge, MA, pp. 47-66.

Delgado, M.M, and Bottoms, S. (2010), "Investigación sobre teatro y artes escénicas en el Reino Unido", *Cairon*, Vol. 13 No. 1, pp. 153-187.

Didegah, F., and Thelwall, M. (2013), "Which factors help authors produce the highest impact research? Collaboration, journal and document properties", *Journal of Informetrics*, Vol. 7 No. 4, pp. 861-873.

ERA (2015), "ERA Discipline Matrix", http://www.arc.gov.au/era/era_2015/2015_keydocs.htm (accessed 13 February 2015)

Franceschet, M., and Costantini, A. (2011), "The first Italian research assessment exercise: A bibliometric perspective", *Journal of Informetrics*, Vol. 5 No. 2, pp. 275-291.

Garfield, E. (1979), Citation *indexing: Its theory and application in science, technology, and humanities*, Wiley, New York.

Glänzel, W. (2002), "Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies", *Library Trends*, Vol. 50 No. 3, pp. 461-473.

Gussak, D. (2006), "Effects of art therapy with prison inmates: A follow-up study", The Arts in Psychotherapy, Vol. 33 No. 3, pp. 188-198.
Heaney, C. J. (1992), "Evaluation of music therapy and other treatment modalities by adult psychiatric inpatients", *Journal of Music Therapy*, Vol. 29 No. 2, pp. 70-86.
HEFCE (2009), "Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework", http://www.hefce.ac.uk/pubs/year/2009/200939/name,63842,en.htmlhttp://www.hefce.ac.uk/pubs/year/2011/201103/name,63893,en.html (accessed 13 February 2015)
HEFCE (2011), "Analysis of data from the pilot exercise to develop bibliometric indicators for the REF: The effect of using normalised citation scores for particular staff characteristics", http://www.hefce.ac.uk/pubs/year/2011/201103/name,63893,en.html (accessed 13 February 2015)
HEFCE (2014), "Citation data", http://www.ref.ac.uk/about/guidance/citationdata/ (accessed 13 February 2015)
Hellqvist, B. (2010), "Referencing in the humanities and its implications for citation analysis", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 2, pp. 310-318.
Heritage, P. (2005), "Parallel power: Shakespeare, gunfire and silence", *Contemporary Theatre Review*, Vol. 15 No. 4, pp. 392-405.
Holden, J. (2004), "Creating cultural value: how culture has become a tool of government policy", Demos, http://www.demos.co.uk/files/CapturingCulturalValue.pdf (accessed 13 February 2015)
Hudson, R. (2007), "How linguistics has influenced schools in England", *Language and Linguistics Compass*, Vol. 1 No. 4, pp. 227-242.
Hyland, K. (2004), *Disciplinary discourses: Social interactions in academic writing*, University of Michigan Press, Ann Arbor, MI.
Ioannidis, J. P. (2005), "Contradicted and initially stronger effects in highly cited clinical research", Jama, Vol. 294 No. 2, pp. 218-228.
Johnson, L. M. (2008), "A place for art in prison: Art as a tool for rehabilitation and management", *Southwest Journal of Criminal Justice*, Vol. 5 No. 2, pp. 100-120.
Kontos, P. C., and Naglie, G. (2007), "Expressions of personhood in Alzheimer's disease: An evaluation of research-based theatre as a pedagogical tool", *Qualitative Health Research*, Vol. 17 No. 6, pp. 799-811.
Korevaar, J. C. and Moed, H. (1996), "Validation of bibliometric indicators in the field of mathematics", *Scientometrics*, Vol. 37 No. 1, pp. 117-130.
Kousha, K. and Thelwall, M. (in press), "Can Amazon.com reviews help to assess the wider impacts of books?", *Journal of the Association for Information Science and Technology*.
Lariviere, V., Gingras, Y., and Archambault, É. (2006), "Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities", *Scientometrics*, Vol. 68 No. 3, pp. 519-533.
MacRoberts, M. H., and MacRoberts, B. R. (1996), "Problems of citation analysis", Scientometrics, Vol. 36 No. 3, pp. 435-444.
Merton, R. K. (1973), *The sociology of science: Theoretical and empirical investigations*, University of Chicago Press, Chicago.
Moed, H. F. (2006), *Citation analysis in research evaluation*, Springer, Berlin.

Moed, H. F. (2008), "UK Research Assessment Exercises: Informed judgments on research quality or quantity?", *Scientometrics*, Vol. 74 No. 1, pp. 153-161.

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics*, Vol. 97 No. 3, pp. 767-777.

Nederhof, A. J., and Van Raan, A. F. (1993), "A bibliometric analysis of six economics research groups: A comparison with peer review," *Research Policy*, Vol. 22 No. 4, pp. 353-368.

Nelson, R. (2013), *Practice as Research in the Arts: Principles, Protocols, Pedagogies, Resistances*, Palgrave Macmillan, Houndmills.

Norris, M., and Oppenheim, C. (2003), "Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE", *Journal of Documentation*, Vol. 59 No. 6, pp. 709-730.

Nussbaum, M. C. (2012), *Not for profit: Why democracy needs the humanities*, Princeton University Press, Princeton NJ.

Ochsner, M., Hug, S. E., and Daniel, H. D. (2014), "Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies", *Zeitschrift für Erziehungswissenschaft*, Vol. 17 No. 6, pp. 111-132.

Oldfield, A., (2006), *Interactive Music Therapy in Child and Family Psychiatry Clinical Practice, Research and Teaching*, Jessica Kingsley, London.

Oppenheim, C., and Summers, M. A. (2008), "Citation counts and the Research Assessment Exercise, part VI: Unit of assessment 67 (music)", *Information Research: An International Electronic Journal*, Vol. 13 No. 2, http://www.informationr.net/ir/13-2/paper342.html

Oppenheim, C. (1995), "The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science university departments", *Journal of Documentation*, Vol. 51 No. 1, pp. 18-27.

Oppenheim, C. (1997), "The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology", *Journal of Documentation*, Vol. 53 No. 5, pp. 477-487.

PBRF (2013), "Quality Evaluation Guidelines 2012", http://www.tec.govt.nz/Documents/Publications/PBRF-Quality-Evaluation-Guidelines-2012.pdf (accessed 13 February 2015)

Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010), Altmetrics: A manifesto. http://altmetrics.org

REF2014 (2012a). Assessment framework and guidance on submissions. http://www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

REF2014 (2012b), "Panel criteria and working methods. Main Panel D criteria", http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12_2D.pdf (accessed 13 February 2015)

REF2014 (2015a), "Research Excellence Framework 2014: Overview Report by Main Panel A and Sub-panels 1-6", http://www.ref.ac.uk/panels/paneloverviewreports/ (accessed 13 February 2015)

REF2014 (2015b), "Research Excellence Framework 2014: Overview Report by Main Panel B and Sub-panels 7-15", http://www.ref.ac.uk/panels/paneloverviewreports/ (accessed 13 February 2015)

REF2014 (2015c), "Research Excellence Framework 2014: Overview Report by Main Panel C and Sub-panels 16-26", http://www.ref.ac.uk/panels/paneloverviewreports/ (accessed 13 February 2015)

REF2014 (2015d), "Research Excellence Framework 2014: Overview Report by Main Panel D and Sub-panels 27-36", http://www.ref.ac.uk/panels/paneloverviewreports/ (accessed 13 February 2015)

Rinia, E. J., Van Leeuwen, T. N., Van Vuren, H. G., and Van Raan, A. F. (1998), "Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands", *Research Policy*, Vol. 27 No. 1, pp. 95-107.

Renfrew, K. and Green, H. (2014), "Support for Arts and Humanities Researchers Post PhD. British Academy and The Arts and Humanities Research Council", http://www.ahrc.ac.uk/News-and-Events/News/Documents/Support%20for%20Arts%20and%20Humanities%20Researchers%20Post-PhD.pdf (accessed 13 February 2015)

Rothman, J. (2014), "Fixing the PhD", *The New Yorker*, 4 June. http://www.newyorker.com/books/joshua-rothman/fixing-the-ph-d (accessed 13 February 2015)

Seglen, P. O. (1998), "Citation rates and journal impact factors are not suitable for evaluation of research", *Acta Orthopaedica*, Vol. 69 No. 3, pp. 224-229.

Small, H. (2013), *The value of the humanities*, Oxford University Press, Oxford.

Smith, A. T., and Eysenck, M. (2002), "The correlation between RAE ratings and citation counts in psychology", http://cogprints.org/2749/

Stige, B. (2005), "Research as practice", *Nordic Journal of Music Therapy*, Vol. 14 No. 2, pp. 90-90.

Thelwall, M. and Sud, P. (2014), "No citation advantage for monograph-based collaborations?", *Journal of Informetrics*, Vol. 8 No. 1, pp. 276-283.

Thomas, P. R., and Watkins, D. S. (1998), "Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications", *Scientometrics*, Vol. 41 No. 3, pp. 335-355.

Thompson, J. (2009), *Performance Affects: Applied Theatre and the End of Effect,* Palgrave Macmillan, Basingstoke.

Torres-Salinas, D., Robinson-Garcia, N., Campanario, J. M., and López-Cózar, E. D. (2014), "Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index", *Online Information Review*, Vol. 38 No. 1, pp. 24-42.

Vacca, J. S. (2004), "Educated prisoners are less likely to return to prison", *Journal of Correctional Education*, Vol. 55 No. 4, pp. 297-305.

van Raan, A. F. J. (1998), "In matters of quantitative studies of science: The fault of theorists is offering too little and asking too much", *Scientometrics*, Vol. 43 No. 1, pp. 129-139.

Wainer, J., and Vieira, P. (2013), "Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of Brazilian scientists", *Scientometrics*, Vol. 96 No. 2, pp. 395-410.

Warner, J. (2000), "A critical review of the application of citation studies to the Research Assessment Exercises", *Journal of Information Science*, Vol. 26 No. 6, pp. 453-459.

Wolffe, J. (2011), "Why religious history matters: Perspectives from 1851", in Bate, J. (Ed.), *The public value of the humanities*, Bloomsbury Academic, London, pp. 44-55.

Appendix: Tests of correlations between peer review scores and citation metrics

Despite the above limitations of citation analysis for the humanities, it may still be reasonable to use citation counts if they could be shown to correlate with expert judgements of research quality or impact, even though the mechanism through which this occurred was opaque. This section reviews research that has attempted to assess the strength of correlation between peer review judgements and citation-based indicators across different disciplines to give context to correlations for arts and humanities fields.

Citation-based indicators must be normalised for year of publication and, if compared between different fields, must also be field normalised in some way (Aksnes and Taxt, 2004). These steps would reduce the influence of two of the largest sources of citation bias. Most studies that have compared citations to human judgements have normalised the citation data to some extent or have at least split it up into broad fields and have analysed articles from only a relatively small number of years. After normalisation, the strength of the correlation between the citation metric and human scores indicates how closely the two would give similar outcomes. For convenience the Dancy and Reidy (2004) naming convention for correlations will be used here: Strong (0.7-0.9); moderate (0.4-0.6); and weak (0.1-0.3). In the discussion below, comparisons between correlations are approximate due to the different normalisation methods and sets of articles used in each case. Although the existence of a statistically significant positive correlation between citation counts and peer judgements indicates that citations indicates that better research tends to be more cited, the strength of the correlation depends to a large extent on the degree of aggregation, the citation window used, and the range of subject areas compared. For example, using citation counts from different citation windows (e.g., taking all citations to date to a set of articles from 2 or more different years) will tend to reduce correlations because older articles will tend to have more citations than younger articles, irrespective of quality. Similarly, comparing articles from different fields will tend to reduce the correlation because articles in fields with high citation norms will tend to be more cited than other articles, irrespective of quality. Finally, aggregating articles together before calculating a correlation coefficient should dramatically increase the correlation. For example, the correlation between the peer review scores for individual articles and their citation counts is 0.1 then the correlation between the average peer review scores of a set of departments (e.g., 50 departments, each based on the average peer scores and citation counts for about 500 papers) could easily be above 0.9 – assuming that there are no systematic biases, such as some departments specialising in particularly high (e.g., pure research) or low citation areas of research (e.g., applied research). The averaging process is complicated by the skewed nature of citations, however, because individual extremely highly cited papers can make a substantial difference to a departmental average.

At the individual article level, there are few studies that compare citations to peer judgements because most research assessment exercises only judge groups of researchers or do not publish their evaluations of individual articles, even if they make them (e.g., the UK Research Assessment Exercise [RAE] and REF). Nevertheless, one study has shown that articles chosen by mathematics as top in their field tend to be much more highly cited than comparable mathematics articles

(Korevaar and Moed, 1996), although this does not prove that a similar relationship would hold for more typical articles.

For individual authors, one study has compared bibliometric measures for Brazilian scientists by field with a career decision about whether their scholarship (promoted, demoted, maintained) (Wainer and Vieira, 2013). There was a weak correlation overall (0.2) between average Scopus citations per paper and scholarship review outcomes, varying from astronomy (-0.7) to mechanical engineering (0.6). Brazilian researchers may not focus on international journals, however, and the numbers involved were not large: under 200 researchers in most categories. Moreover, absolute numbers of publications would also have been taken into account in the career evaluations.

For groups of researchers within an area, such physics departments, correlations should be high as long as the groups are narrowly defined in terms of discipline. A comparison of citations to groups of Italian researchers, based articles from 2001-2003, with peer review scores used wide categories that combined similar fields and a variable citation window, but still found mostly strong or moderate correlations (Franceschet and Costantini, 2011). In the medical research area the correlations were moderate in the medical sciences (0.6) and agricultural sciences and veterinary medicine (0.5) when averaged across research units. In the natural sciences, the correlations were strong in physics (0.8), earth sciences (0.8), and biology (0.7), and moderate in chemistry (0.6). In engineering, the correlations were moderate in industrial and information engineering (0.6) and mathematics and computer sciences (0.5) and weak in civil engineering and architecture (0.3). In the social sciences, the correlations were strong in economics and statistics (0.4).

For 56 condensed matter physics research groups in the Netherlands, peer review judgements of each group correlated moderately (0.6) with the average number of citations per publication produced by the group (Rinia, Van Leeuwen, Van Vuren, and Van Raan, 1998). A comparison citation data and peer review scores for six economics research groups found that the results were complementary and therefore that the combination of the two would give the best results (Nederhof and Van Raan, 1993).

Many studies have compared average citations to the research of UK departments with their Research Assessment Exercise (RAE) rankings, which are based on peer judgements of a limited set of outputs (e.g., four per academic submitted for assessment). There has been an unavoidable mismatch in the data compared in these studies but most have calculated the average citations per published article as one of the metrics, although it would also be reasonable to calculate the average per member of staff. In psychology departments the rankings for 1996 and 2001 correlate strongly with the average number of citations per member of staff based on publications from 1998 (Smith and Eysenck, 2002). RAE 2001 scores for UK archaeology departments correlate strongly (0.7) with average citations to the publications of staff in the department (Norris and Oppenheim, 2003). Statistically significant positive correlations have also been found between average citations and 1992 RAE scores for departments in genetics, anatomy, archaeology (Oppenheim, 1997) and library and information science (Oppenheim, 1995), for 1996 scores in Business and Management (Thomas and Watkins, 1998) and for 2001 scores in music (0.8) (Oppenheim and Summers, 2008). Peer evaluations of different models for calculating bibliometric indicators found that average citation scores per department within specific subject areas worked best if only the best papers were included rather than all papers produced, and considered the results to

be "credible", but with significant discrepancies, in medicine, biological and physical sciences, and psychology, but not in social sciences, mathematics, health sciences, engineering and computer science (HEFCE, 2009). One study has compared average (field/journal and year) normalised citation counts to RAE articles with the average RAE scores of their publishing groups in six different fields of study. The Spearman correlations were: biology 0.53; chemistry 0.62; physics 0.53; mechanical, aeronautical and manufacturing engineering 0.18; geography and environmental studies 0.47; sociology 0.47; history 0.38 (Mryglod, Kenna, Holovatch, and Berche, 2013). This seems to be the most methodologically sound of the RAE-related studies and suggests that, whilst there is a correlation between appropriately normalised average citation counts and average quality at the level of the humanities research group, it is likely to be lower than in all other areas of scholarship, except perhaps for engineering.

A study of peer ratings of the overall "quality of graduate faculty" in comparison to a citation-related indicator for schools ten fields for universities in the USA, found statistically significant positive correlations in all cases, from Developmental Biology (0.3) and Zoology (0.3) to Mathematics (0.8), with one social science, Psychology (0.7). The results suggested that peer judgements of departments were influenced by both the size of the department and the quality of its research, at least as reflected by the average citation impact of its articles (Anderson, Narin, and McAllister, 1978).

In summary, whilst most of the evidence of a relationship between appropriately normalised average citation scores suggest that they tend to correlate positively with peer judgements, these correlations vary by discipline. Although the correlations are stronger at the level of research groups than for individual researchers or papers, they are not high enough to replace important peer judgements. Nevertheless, the existence of positive correlations in some cases, even in subjects like maths, music and archaeology, suggests that there might be ways in which they could be used to inform peer judgements, even in the formal sciences, arts and humanities.