When rating systems do not rate

Evaluating ERA's performance

Paul Henman
University of Queensland

Scott D Brown & Simon Dennis
University of Newcastle

In 2015, the Australian Government's Excellence in Research for Australia (ERA) assessment of research quality declined to rate 1.5 per cent of submissions from universities. The public debate focused on practices of gaming or 'coding errors' within university submissions as the reason for this outcome. The issue was about the in/appropriate allocation of research activities to Fields of Research. This paper argues that such practices are only part of the explanation. With the support of statistical modelling, unrated outcomes are shown to have also arisen from particular evaluation practices within the discipline of Psychology and the associated Medical and Health Sciences Research Evaluation Committee. Given the high stakes nature of unrated outcomes and that the evaluation process breaches public administration principles by being not appealable nor appropriately transparent, the paper concludes with recommendations for the strengthening ERA policy and procedures to enhance trust in future ERA processes.

Keywords: higher education policy, universities, research performance, performance assessment, performance measurement, Excellence in Research for Australia, ERA, Australian Research Council, ARC, public administration

Introduction

On 4 December 2015, the Australian Research Council (ARC) released the results of its third Excellence in Research for Australia (ERA) assessment process. The process assessed disciplines, or Fields of Research (FoR), in each of 41 Australian universities, based on their own submissions. Each submission, or Unit of Evaluation (UoE), consisted of research indicators including research funding, research publications, applied indicators, and esteem measures. Each discipline at each university that made a submission was rated from 1 to 5, being 'well below world standard' to being 'well above world standard'. Where institutions did not make a submission for a particular discipline, a category of 'not assessed' (n/a) was given.

The ERA process has been subjected to academic analysis and commentary (Gable, 2013), as has the wider and growing practice of national research assessment exercises, such as the UK's Research Excellence Framework (REF) and its predecessor, the Research Assessment Exercise (RAE) (Elton, 2000; Geuna & Martin, 2003). Many of these articles are highly critical of these forms of evaluation and their authors have argued that they do not achieve their stated goals, and also lead to dysfunctional managerial control over traditional academic practices (Bonnell, 2016; Kwok, 2013; Martin, 2011). While many of these articles challenge the performance metric

approach, many more engage with the contours and specifics of performance evaluation design. For example, the ERA 2010 allocation of a four-scale ranking to journals was subjected to such widespread challenge due to its subjective and political construction (Haslam & Koval, 2010; Vanclay, 2011) that it was finally dropped.

In ERA 2015 a new twist to the evaluation process occurred. In addition to the five-scale rating, a new category was created, 'not rated' (n/r). This new outcome was given for disciplinary areas within an institution that made a submission, but 'due to coding errors' ERA committees were unable or unwilling to allocate it a numerical rating from one to five (ERA, 2015, p. 364). The failure of the rating system to rate some submissions has generated considerable public and private debate, including media coverage.

Unfortunately, the debate so far has shed little light on how submissions came to be categorised as n/r, and speaks to a wider discomfort in the university sector about the black-box, opaque nature of ERA rating exercises. Indeed, there is considerable secrecy in the process. University submissions are confidential, and members of the Research Evaluation Committees (RECs) sign strict confidentiality agreements to ensure the confidentiality of submissions to the rating process and of the process itself. While the members of the RECs are known (http://www.arc.gov.au/era-2015-research-evaluation-committee-rec-members) the reviewers who provide input into peer-reviewed disciplines are not, and there is no process for appealing the rating results. Institutions can, however, appeal the process by which ratings were made if they are viewed as infringing the stated process, but as the process is largely hidden it would be hard to make a case on those grounds.

There are significant public interest grounds for analysing this situation. Firstly, there are important public administration principles at stake, including due process, administrative fairness, transparent and accountable public administration, and decision making being subject to appeal (Bovens *et al.*, 2014; Weber, 2009, pp. 196-244). Secondly, university reputations are enhanced or degraded based on ratings, and the not rated outcomes have cast significant doubt about the propriety of those seven institutions who received them. Thirdly, such universities also need clear feedback to identify perceived problems in their submissions in order to avoid future n/r outcomes.

In an effort to better understand the administrative black box, we examined the ERA production of n/r UoE using both textual analysis of public reporting and statistical examination of the outcomes. One of our goals was to further investigate and understand the locus of the problems associated with n/r outcomes. Furthermore, our analyses provide a basis for suggestions for enhancing the public governance of ERA processes to ensure they maintain confidence and credibility for future ERA rounds, including its expansion in 2018 to cover research impact.

The next section presents an overview of the n/r outcomes with respect to institutions, FoR and ERA RECs. These data and their patterns are then considered in the light of public statements about the n/r UoE, and statistical analysis using Bayesian statistical modelling to assess the probabilities of the n/r outcomes. These observations are interpreted to identify the best explanation for the n/r outcomes. The concluding section considers what these findings suggest for the possible changes to ERA rules and administrative processes.

... university reputations are enhanced or degraded based on ratings, and the not rated outcomes have cast significant doubt about the propriety of those seven institutions who received them.

ERA 2015's not rated submissions

For ERA 2015, each of 41 universities was invited to make submissions for rating in any of 179 FoR. These fields are defined by the FoR codes in the Australian and New Zealand Standard Research Classification (ABS, 2008). Here, and throughout, we exclude from analysis the two-digit FoR codes, and focus on the more fine-grained divisions of the four-digit codes, of which there are 157. Together, the 41 universities made 1,802 submissions. Of these, only 27, or 1.5 per cent, were categorised as n/r. These 27 n/r UoE were distributed across seven of the 41 (17 per cent) participating universities (see Table 1). University of Wollongong had the most, with 13 out of their 54 (24 per cent) submissions n/r, Victoria University second with 8 of 29 (28 per cent) submissions n/r, while the remaining five had only one or two n/r submissions. Data on ERA ratings have been obtained from the ARC website (www.arc.gov.au/era-outcomes) and reports (ARC 2015).

Another way to view the n/r units is by discipline (or Field of Research – FoR), instead of institution. Of the 157 different disciplines defined by the FoR codes, 20 disciplines generated at least one n/r unit of evaluation (see Table 2). Only two of the FoR codes yielded a n/r

Table 1: Unrated (n/r) submissions by institution

Institution	*Number of n/r from total of institution's UoE*	*Percentage n/r of institution's UoE*
University of Wollongong	13 of 54	24%
Victoria University	8 of 29	28%
University of Tasmania	2 of 53	3.8%
Central Queensland University	1 of 14	7.1%
Edith Cowan University	1 of 27	3.7%
RMIT University	1 of 39	2.6%
University of Newcastle	1 of 57	1.7%

Source: www.arc.gov.au/era-outcomes

Table 2: Unrated (n/r) submissions by Field of Research

Field of Research		*Number of unrated (n/r) submissions*
0103	Numerical and Computational Mathematics	1
0204	Condensed Matter Physics	1
0601	Biochemistry and Cell Biology	1
0904	Chemical Engineering	1
0905	Civil Engineering	1
0908	Food Sciences	1
0915	Interdisciplinary Engineering	1
0999	Other Engineering	1
1103	Clinical Sciences	1
1111	Nutrition and Dietetics	1
1115	Pharmacology and Pharmaceutical Sciences	1
1116	Medical Physiology	1
1117	Public Health and Health Services	2
1402	Applied Economics	1
1501	Accounting, Auditing and Accountability	1
1503	Business and Management	1
1505	Marketing	1
1599	Other Commerce, Management, Tourism and Services	1
1701	Psychology	5
1799	Other Psychology and Cognitive Sciences	1

Source: www.arc.gov.au/era-outcomes

Table 3: Unrated (n/r) submissions by Research Evaluation Committee (REC)

Research Evaluation Committee	*Number of unrated (n/r) from total evaluated*	*Percentage unrated of submissions considered*
Biological and Biotechnological Sciences (BB)	3 of 199	1.5%
Economics and Commerce (EC)	5 of 171	2.9%
Education and Human Society (EHS)	0 of 214	0%.
Engineering and Environmental Sciences (EE)	5 of 214	2.3%
Humanities and Creative Arts (HCA)	0 of 308	0%
Mathematical, Information and Computing Sciences (MIC)	1 of 153	0.7%
Medical and Health Sciences (MHS)	12 of 321	3.7%
Physical, Chemical and Earth Sciences (PCE)	1 of 222	0.5%

Source: www.arc.gov.au/era-outcomes

category for more than one university: Public Health and Health Services (FoR code 1117) and Psychology (FoR code 1701), which yielded two and five n/r outcomes, respectively.

A third way to view these results is by the committees that allocated ratings. The ERA ratings were determined by the eight RECs, each of which was responsible for rating submissions in a different subset of FoR codes. The distribution of n/r outcomes across RECs is shown in Table 3. Just one REC, the Medical panel, was responsible for almost half of all n/r outcomes from ERA 2015 (12 outcomes, or 3.7 per cent).

Analysing the reasons for not rated outcomes

One obvious interpretation of the pattern of these outcomes is that the problems of n/r submissions lie within universities, and in particular, what they submitted. Indeed, in the context of high-stakes performance measurement, gaming is to be expected (Bevan & Hood, 2006; Hood, 2006; Jacob, 2005), and rules are likely to be bent as far as possible to enhance institutional performance, as indicated by the ratings. This perspective

is certainly what has been implied and is a widespread perception within the sector, as evidenced in the mass media articles cited below.

Analysing not rated using public statements

Prior to the release of the ERA 2015 results, a story in the Fairfax media reported that 'Several universities are being threatened with tough penalties for allegedly providing data that would artificially boost their performance on [ERA]' (Knott, 2015, 18 November), and referred specifically to University of Tasmania and Central Queensland University. The ARC's CEO Professor Aidan Byrne responded that, '…it is not correct that either of the universities named "coded" journal articles "multiple times" to "inflate a university's results"'. He proceeded to say that REC panels had raised queries about data and that such queries 'account for well under 2 per cent of the UoEs submitted for assessment', and that such assessment will be 'based on clear and robust processes with the rules of submission clearly stated when released in July 2014'. The media's presentation clearly constructs some universities as gaming the system with alleged 'data manipulation' that will be countered by clear, fair and robust review processes. More directly, this coverage suggests that some institutions may have breached the ERA rules.

Following the release of the ERA results, *The Australian* summarised the ERA n/r outcomes by institution (Loussikian, 2015, 9 December). The newspaper article referred to '[A] significant number of coding errors in submissions' as the cause of the n/r outcomes, and the ARC is reported to have referred to 'coding issue[s]' (see also ARC, 2015, p. 364), statements that reinforce the view that the ERA rules were breached. However, the University of Wollongong – the institution with the most n/r outcomes – was reported to state that 'UoW's ERA 2015 submission was *prepared in accordance with the…submission guidelines* and followed the same process as previous submissions' (ARC, 2015, p. 364, emphasis added), thereby countering the view that ERA rules were breached.

Subsequently, Victoria University contracted Emeritus Professor Alan Lawson from the University of Queensland to undertake a review of their ERA submission process in the light of their high proportion of n/r submissions. According to a report in *The Australian* (Loussikian, 2016, 16 March), Professor Lawson found that the university allocated some research performance indicators (e.g. publications) to FoRs not related to that indicator, apparently in breach of the ERA rules that outputs can only be assigned to a specific field of research if 'they are relevant to that output'.

To be sure, the ERA rules state 'FoRs should be assigned to an output if they are relevant to that output' (ARC, 2014, p. 32). Indeed, for each journal and conference proceedings, the ARC assigns up to three FoR codes, and institutions must submit journal or conference publications to these FoR codes. The exception is the 'reassignment rule' which states that another FoR code can be used when 'publications which have significant content (66 per cent or more) that could best be described by [that] four-digit FoR code' (ARC, 2014, p. 33). The allocation of books, book chapters and research funding does not have this detail, but remains subject to the overarching rule stated at the start of this paragraph.

Notably, under the rules, the FoR codes assigned to researchers have no relationship to the FoR code assigned to their publications. Rather, the ERA rules state 'FoR assignment should describe the focus of the activities of the researcher' (ARC, 2014, p. 29). This results in situations such as an engineer publishing in an international development journal about engineering projects in developing countries, where the FoR codes assigned to the researcher may relate entirely to engineering (FoR two digit code 09), and yet the journal could be assigned by the ARC or the institution to FoR codes in Economics (FoR two digit code 14) or Studies in Human Society (FoR two digit code 16).

Returning to the case of Victoria University, the reporting of Professor Lawson's review can be interpreted as implying that Victoria University's submission did not pass the overarching ERA research output rule, and instead used 'complex computer models' without overarching 'academic judgement' to allocate FoR codes to outputs on a numerical basis, thereby separating publication content from its most relevant FoR code.

Important questions remain. Are 'problems' or gaming within institutions the sole explanation for the n/r ERA outcomes? Is failure of universities to provide appropriate submissions that accord with the 'clear and robust' ERA processes with 'independent' committees with 'integrity' the only interpretation or full explanation of a failure of rating system to rate? Is it just the fault of institutions who have not played by the rules? Alternatively, are there public administration insights about the ERA rules and processes that also need to be considered? Indeed, the ARC reported that the reason for n/r UoEs is 'not the same reason for each unit of [evaluation] that has received this rating' (Loussikian, 2015, 9 December). Accordingly, it is important to identify whether the patterns of n/r UoEs reflect problems within the ERA system beyond the university level. To provide further

insight into these questions, we undertook statistical modelling to consider these questions.

Analysing not rated patterns statistically

The distribution of n/r outcomes is uneven across institutions (see Table 1), which reinforces the interpretation that erroneous institutional submissions explain the n/r submissions. However, the distribution of n/r outcomes also appears uneven across FoR codes (Table 2), and across RECs (Table 3). This variability raises questions about whether the greater number of n/r outcomes in certain categories is more likely to be caused by differences between institutions, differences between the RECs, differences in the ways that FoR codes were evaluated, or a combination of each. Put another way, while the Victoria University case demonstrates that there can be systemic practices with ERA submission processes within a specific institution, is it also possible that there may be practices within specific disciplines or specific REC processes that can explain the patterns of n/r submissions?

There are standard and well-accepted statistical approaches for addressing these questions. Similar questions arise in educational and psychological testing settings, for example when many students take an exam consisting of many questions, and the examiners wish to know whether any students performed particularly well or poorly, and also whether questions were particularly difficult or easy.

We used a Bayesian approach based on nested model tests (Congdon, 2006) to identify whether the proportion of n/r outcomes was unusual for any particular institution, FoR code, or REC. In other words, the tests can identify whether there are differences between different institutions, different FoR codes or different RECs in the presence of n/r outcomes that are not random statistical variations. The analysis approach for institutions, FoR codes and RECs was identical in each case, which we outline here for the institution-focussed analysis. We first fitted a null model to the data, which treats all institutions identically, and estimated a single parameter, representing the probability of a n/r outcome. All UoEs, from all institutions and all FoR codes, were assumed to have this same probability of (independently) yielding a n/r outcome, which leads to a binomial distribution for the number of n/r outcomes. Next, we examined an 'outlier' model, which tested whether one particular institution was different from the others. For this model, we chose one institution and estimated one probability for that institution, and a different probability for all other combined institutions. We then used Bayes factors to compare which of these two models – the null model or the outlier model – provided the best description of the full data set. In all cases, we used the same prior distribution for the probability of yielding a n/r outcome. This prior distribution was Beta(11,1), which can be interpreted as a prior expectation consistent with previously observing ten ERA outcomes, none of which were n/r. We tested many other prior distributions, from uniform, Beta(1,1), up to Beta(41,1), and the results were qualitatively unchanged.

The null model against which all others were compared showed that the data favoured an overall probability of n/r outcomes around 1.5 per cent (this was the median of the posterior distribution), with a 95 per cent highest posterior density (HPD) interval from 1.03 per cent to 2.16 per cent. Note that 1.5 per cent was chosen as this is the percentage of UoEs that were n/r.

Institutions as outliers: We compared the null model against 41 models, each one of which treated one particular institution as having a different rate of n/r outcomes from all the others. Figure 1 summarises results from these models. The 41 institutions are arrayed along the x-axis, and above each one the black error bar shows the 95 per cent HPD interval for probability of a n/r outcome from that institution (the black circle is the posterior median). The shaded rectangle covers the 95% HPD for the probability of a n/r outcome from the other 40 institutions combined. In most cases, the black error bars overlap the red rectangles which means that there is no evidence that the singled-out institution has a different rate of n/r outcomes from the others. For just two institutions (University of Wollongong, and Victoria University) there was evidence in favour of a different rate of n/r outcomes than the others. To quantify these results, we calculated Bayes factors to quantify the evidence in favour of each outlier model over the null model. These are printed at the top of Figure 1. In almost all cases, the Bayes factors are less than 1, indicating evidence in favour of the null model, namely that allocation of n/r is the same in this institution as the others. Only for the two institutions mentioned above do the Bayes factors indicate strong evidence in favour of a different rate of n/r outcome. Put differently, these findings suggest that there is something different about these two institutions' ERA outcomes (and thus their submissions) compared with other institutions, and that there is no discernible difference between the other institutions' ERA outcomes (and thus their submissions).

FoR Codes as Outliers: Our second analysis investigated the evidence for each FoR code yielding a different rate

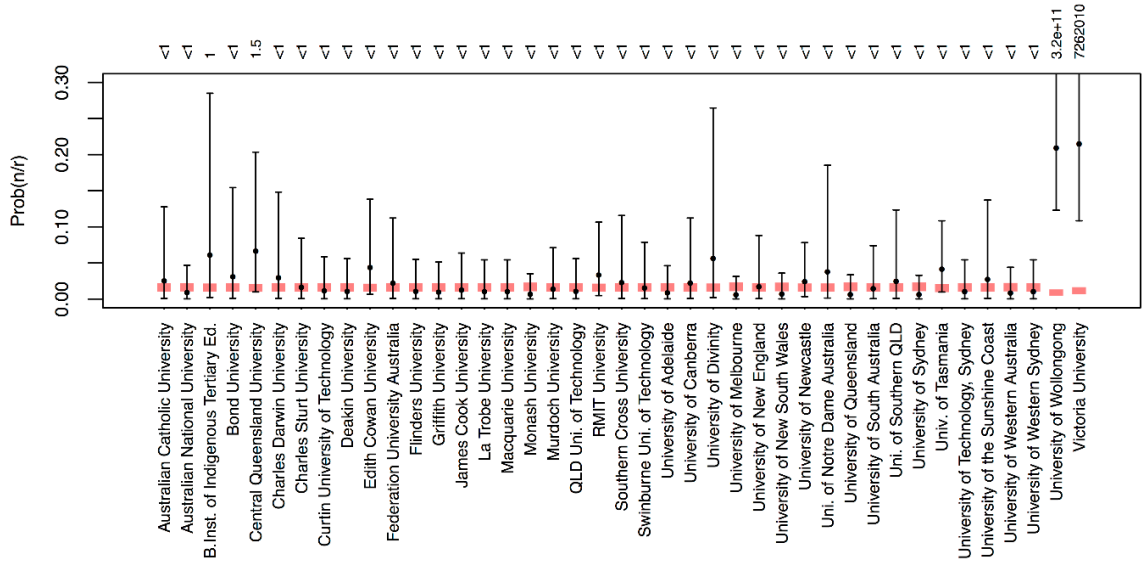

Figure 1: Bayes factors in favour of an outlier institution

of n/r outcomes than the others. To simplify, we grouped all the four-digit FoR codes into their two-digit categories (e.g. the FoR=18 analysis grouped together FoRs 1801, 1802, and 1899). Figure 2 summarises this analysis, using the same format as Figure 1. Once again, most FoR codes showed evidence in favour of having the same rate of n/r outcomes as the others. Only one FoR code was different – FoR=17, Psychology. This code had a posterior median probability of yielding a n/r result nearly an order of magnitude higher than the others, at 13.3 per cent (HPD: 5.8%-24.3%). The Bayes factor in favour of the outlier model which treated FoR=17 as different from the others indicated strong evidence for this over the null model: 3,909-to-1. Putting these findings another way, they suggest that there is something different about the ERA outcomes for Psychology FoR code as against other FoR codes, and that there is no discernible difference between the other FoR codes' ERA outcomes.

RECs as Outliers: The final analysis investigated evidence for each of the eight RECs being different from the others (see Figure 3). There was some evidence that the Humanities REC statistically provided fewer

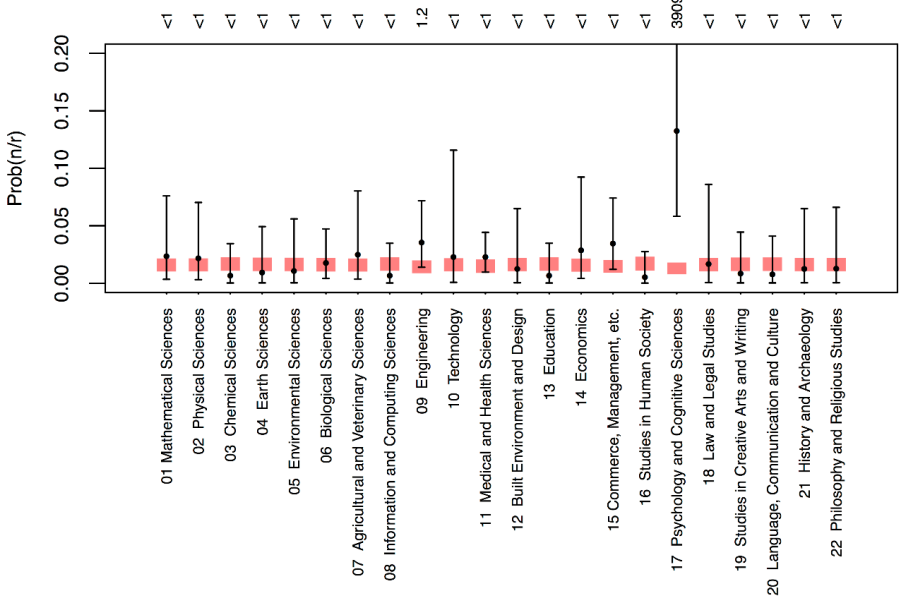

Figure 2: Bayes factors in favour of an outlier FoR Code

n/r outcomes than other panels, but the evidence was not strong (6.6-to-1). There was stronger statistical evidence that the Medical REC gave more n/r outcomes than other panels (34-to-1). Putting these findings another way, they suggest that there is something different about the ERA outcomes allocated by the Humanities and Medical RECs than other RECs, and that there is no discernible difference between the other RECs' ERA ratings. Importantly, the Humanities REC is found to be statistically less likely to give n/r outcomes (indeed, they provided none), while the Medical REC is found to be statistically more likely to give n/r. In other words, these modelling results suggest that there is something about the processes within the Humanities REC that means it is less likely to allocate n/r to UoEs than other RECs, and that the Medical REC processes make it more likely to allocate n/r to UoEs. This raises questions about what it is about the composition of and/or processes within these RECs (or the FoRs that they assess) that explains their lesser/greater propensity to generate n/r outcomes.

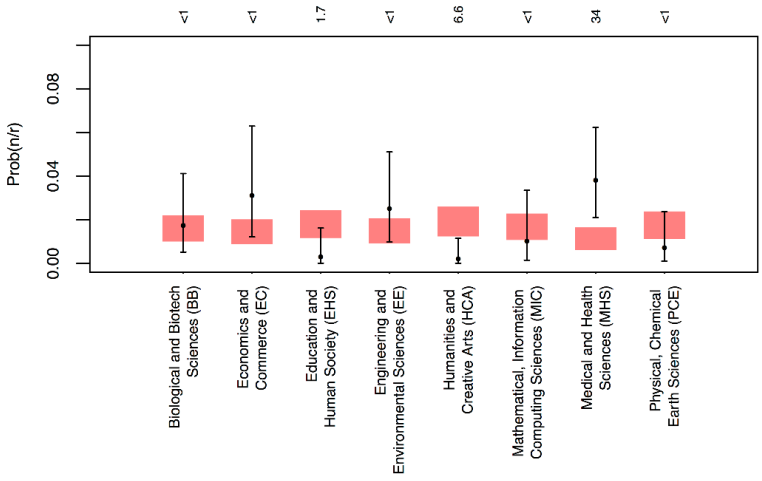

Figure 3: Bayes factors in favour of an outlier REC

Analysing ERA processes

The above findings are notable, as they challenge the suggestion that institutions' submissions were the sole cause of not rated outcomes. Clearly, the ratings produced by the ERA process are based on the submissions from institutions. The fact that there is a statistically unusual allocation of n/r outcomes in Psychology (FoR 17) and in the Medical REC suggests, however, that there is something about the nature of the psychology discipline and/or Medical REC that is different to other disciplines, and RECs that is independent of institutions' submissions. Due to the black box nature of the ERA process, including institutions' submissions to the ERA and the deliberations of the RECs, it is not possible to directly identify what these are. In the absence of such information, we offer some possibilities in the following.

Consider first some possible explanations for why Psychology (FoR 17) outcomes are unlike those for other disciplinary codes. One explanation is that psychology is a broad church. People who identify as psychologists can study anything from the molecular processes that are affected in the progeny of rats that have been stressed, to the neural correlates of predictive process in early audition, to the mathematical properties of diffusion models used to describe simple decisions, to the group processes that underpin conflict in Northern Ireland, to the effectiveness of programs to encourage smoking cessation, to the impact of mindfulness on mental health. Identifying the boundaries of the discipline could be argued to be more difficult than for other disciplines.

There are two aspects of this consideration. The first aspect relates to the types of research in which psychologists are involved. It is very broad. It could be counter-argued that there are similarly other disciplines that contribute to a broad range of research areas – such as statistics (FoR 1604) as a method for countless areas, or engineering (two-unit FoR 09) as a process of building and design in fields diverse as telecommunications, mining, cityscapes, architecture and even social planning, and policy and administration (FoR 1605) relating to all aspects of the natural, social and economic world that governments and societies collectively seek to shape and direct – yet these FoRs did not generate n/r outcomes.

The second aspect relates to the types of research that are regarded as legitimately 'psychological' research concerns, as opposed to have the expertise of psychologists contribute to a research domain. To illustrate more starkly, as statistics is a research method that can be applied almost universally to topics of research, statisticians contribute to research in countless FoRs. However, their contribution to a research domain does not typically contribute to the research field of statistics; it does not develop statistics as a body of knowledge or endeavour further. Statistics acts as a tool to research. It might be argued that what legitimately counts as psychological research (as opposed to the use of psychological expertise in research) is much

wider than what counts as statistical research (as opposed to the use of statistics expertise in research).

These observations in themselves do not directly explain the high likelihood of submissions in psychology recording a n/r outcome than any other field of research. It firstly suggests, however, that the great diversity in what counts as psychological research creates confusion and debate about where the boundary lies between what is and what is not psychology research. Secondly, it might mean that obtaining the corresponding psychological expertise on the REC to cover the breadth could have posed a difficulty for the ARC, resulting in the FoR panel, or REC, not feeling competent to award a rating to some submissions. Put another way, the results suggest that there was a mismatch between the scope that the panel was prepared to admit for the 1701 code and the assumptions that multiple universities employed when compiling their lists of inputs under this code. Recall that there is nothing to suggest here that submissions were in obvious breach of 'the [ERA] rules of submission clearly stated when released in July 2014', but rather that there are differences in opinion relating to the application of the more discretionary elements, notably the 'reassignment rule', whereby journal publications can to be assigned to other FoRs than those allocated by the ARC's ERA 2015 Submission Journal List in certain circumstances.

The above analyses are best guesses that try to interpret the data and processes. Unfortunately, it is not possible to assess the veracity of our analyses given the systemic secrecy of the rating process. The feedback from the ARC to institutions with n/r UoEs could provide further evidence to test our interpretation of the process leading to some n/r outcomes. Indeed, ARC CEO Professor Aidan Bryne explained: 'Where our Research Evaluation Committees (RECs), our independent committees made up of discipline experts nominated by Australian universities, have queried data the ARC has communicated with that university' (ARC, 2015). Such communication is not available to us.

Assessing the quality of ERA assessment processes

We can only speculate as to the basis for n/r outcomes, but it is clear that systematic problems occurred within two universities, and that there are also systemic problems within the rating processes in the Psychology FoR. The existence of rating system shortcomings whose origins lie in the rating processes rather than the universities has significant public administration implications.

The principle of 'administrative fairness', 'administrative justice' or 'due process', which is a hallmark of modern democratic public administrations is multi-faceted (Adler, 2003; Mashaw, 1985). It requires that administrative rules are clear and transparent, that rules are applied equitably and fairly, that decision making processes are open and transparent including explanations for decisions when sought, and that administrative decisions can be appealed (du Gay, 2000; Hunter, 1993; Weber, 2009, pp. 196-244). All these dimensions ensure that public administrators remain accountable and that unfair treatment, bias and corruption do not occur. These classic administrative principles were well understood by Weber over a century ago, and continue today invested in different parlance including 'open government' and 'open data' (Lathrop & Ruma, 2010). In the light of the above findings, it is important to evaluate how well the ERA 2015 processes that led to the overrepresentation n/r outcomes in the Psychology FoR and Medical REC accords with these fundamental public administrative principles. By doing so, possible amendments to the ERA process can be identified.

In relation to the need for clear and transparent administrative rules, the ARC has a comprehensive, published set of rules (ARC, 2014) which also set out the processes of evaluation. The rules provide clear requirements and directions, but also provide a level of flexibility especially as they relate to the allocation of FoR codes to evidence submitted by Universities, such as publications and research funding. This is necessary as the boundaries between disciplines (or FoR) are not clear or rigid. With much contemporary research being conducted by teams of researchers from a variety of disciplines, the discipline/s in which a research project or publication is located can be difficult to specify unambiguously.

It is arguably within this space of discretion that the second principle of applying rules equitably and fairly becomes problematic. It is problematic for institutions in abiding by the rules in applying FoR codes to evaluation evidence, and problematic for ERA evaluators in ensuring they are applied appropriately. To illustrate, consider the 'reassignment rule', which states that another FoR code can be used to those specified in the ERA 2015 Submission Journal List when 'publications which have significant content (66 per cent or more) that could best be described by [that] four-digit FoR code' (ARC, 2014, p. 33). When, for example, might that be, and how might that be determined? By the discipline identity of the authors? By the authors' own sense of the discipline in which the study is located? By the authors' institution's sense of the

discipline in which the study is located? By the journals where the paper is cited? Another potential challenge arises when a publication's authorship is across different institutions. Should a publication's FoR assignment be consistent across institutions, or is it based solely on the perspective of each institution and authors in each institution?

No doubt, the public reporting of Victoria University's submission suggests that Victoria University breached these assignment rules. However, University of Wollongong publicly contends their submission was 'in accordance with the…submission guidelines', and our modelling of n/r outcomes suggests that this would also be the case for all other institutions allocated a n/r outcome. At the same time, it could be surmised from the n/r modelling that reviewers of the Psychology FoR submissions and the Health REC took a different view on compliance with the rules. This is not to suggest that the REC failed to apply rules fairly or equitably, but rather that the rules are not sufficiently robust to achieve agreement between all parties on the relationship between FoR codes and research evidence submitted.

The third administrative principle that decision making processes are open and transparent including explanations for decisions when sought, is clearly not evident in ERA rating exercises. ERA rating processes are largely secret through systematic nondisclosure. REC members and external reviewers (the latter are not publicly named) must sign strict confidentiality agreements and institutions' submissions are not made public despite the bulk of the data – specifically, publications and research funding – being already public at researcher level and aggregate institutional level. Nor are the bases for rating (or not rating) decisions made public. In relation to n/r outcomes, the ARC has publicly stated that relevant institutions have had feedback, however anecdotally it is understood that the feedback to institutions has been minimal. It is also stated in the ERA Submission Guidelines that if the ARC regards part of a UoE as 'incomplete or inaccurate, or contains false or misleading information' it may not submit such information to ERA processes and that the ARC 'will advise the institution of [such] action and provide a statement of reasons' (ARC, 2014: p. 72). These are important matters of procedural justice. For if an institution does not understand the grounds on which they received a n/r outcome, then they are unable to change their organisational processes and submissions for the following ERA. A further implication of not knowing the basis for or process leading to a n/r outcome means that institutions are unable to appeal the decision, a situation that relates to the fourth public administration principle which is that administrative decisions be appealable.

The ERA process largely fails the administrative principle that administrative decisions can be appealed. The scarcity of feedback described above effectively prevents appeals. While the ERA submission guidelines (ARC 2015) makes no mention of appeal rights regarding ERA outcomes, it is understood that institutions have been advised that ERA outcomes are not appealable, only the ERA processes are appealable when those processes do not accord with ERA rules, a situation that institutions are unlikely to know because they are confidential.

Conclusion and consideration of ERA reforms

This paper has analysed, using Bayesian statistical modelling, the loci of n/r outcomes in the ERA 2015 processes. While public statements of such outcomes strongly suggest that n/r outcomes arise from institutional problems, arguably from gaming behaviours, our analyses suggest that this is only part of the explanation. The likelihood of Victoria University and University of Wollongong's n/r outcomes being random is virtually zero. However, there also appears to be systemic differences between the rating processes of the Psychology FoR (1701) and the Medical REC and those of other FoRs and RECs that are not attributable to institutional factors. Specifically, the statistical likelihood of the n/r outcomes in these areas being random is respectively 3900-to-1 and 34-to-1. In relation to this finding, we have proposed that psychology's very diverse research foci combined with the wide discretion given to institutions in attributing research evidence to FoR codes, has led to disagreements between submitting institutions and ERA assessing personnel, resulting in n/r outcomes. This suggests that some changes in ERA policy, procedures and guidelines are necessary to avoid such outcomes in future. By reference to recognised principles of public administration, we further argued that ERA 2015 processes fell short of good public administration.

The essential problem of the ERA process is how to assign research inputs and outputs to FoR codes and to do so in a way that facilitates agreement between the institutions doing the assigning and the REC evaluators. The ARC's approach to assigning FoRs to each journal is necessarily vexed as it is a proxy for the FoR of each article published in that journal. The ARC discovered that assessing journal article quality (or impact) by a proxy of journal quality (or impact factor) was highly problematic

and dropped it. The challenge in this case is more problematic as there needs to be agreement between assigners of FoR to publications and the FoR evaluators of those publications. With this understanding, a number of amendments to the ERA process could be implemented in order to strengthen the ERA process and ensure that institutions can reduce or eliminate the chances of n/r outcomes in future:

Provide Feedback. One straightforward action that could be taken is to provide feedback to each institution about a decision to allocate an n/r outcome. If there are particular publications or research funding that have been deemed inappropriate for a given FoR code, then these should be highlighted to clarify the interpreted scope of the code. This would enable institutions to modify the way in which they allocate publications in future rounds. If there was some other basis for allocating the n/r outcome then that should be articulated. Some action of this kind would seem to be a minimal yet effective response.

Institute an appeals process. Mistakes are sometimes made. Given the high stakes nature of the ERA process, an appeals mechanism would be appropriate to ensure no error has occurred and that RECs are accountable for their decisions.

Remove the 'reassignment rule'. Much of the uncertainty that currently pertains and much of the scope for "gaming" the system occurs as a consequence of the reassignment rule. One option then, would be to remove this provision. Institutions would then be restricted to allocating publications to FoR codes on the basis of the journals in which they appear in the ERA Submission Journal List. Provided they conformed to these rules their submissions would be deemed acceptable. Should this path be adopted, it may be necessary for institutions to be able to interact with the ARC to propose changes to the current assignments, including increasing the maximum number of FoR codes for each publication. If this process occurred prior to the actual submission process then uncertainty would be eliminated and for the non-review based panels, there may be no requirement to meet.

Tighten the 'reassignment rule'. The 'reassignment rule' could to be tightened to limit the scope of reassignments. For example, the rules could require that in cases of publication reassignment to FoR X, it is necessary to have at least one author who has been assigned FoR X. Overall, there could be a higher association between submissions and submitted researchers. For example, currently research income can be submitted under FoR codes not related to a Chief Investigator's FoR, and indeed can be inconsistent with the FoR codes listed on awarded ARC and NHMRC projects. This appears to be a significant anomaly.

A Machine Learning approach. Allowing assignment of publications to FoR codes on the basis of either the journals in which they appear or the FoR assignments of researchers who wrote them would be transparent, but not always entirely adequate. For instance, interdisciplinary journals such as Science or Nature could potentially be assigned to many FoR codes, but one might not want to allow a publication that appears in these journals to be allocated to just any of the available FoR codes. An alternative approach would be to develop a machine classifier capable of taking the journal, title and abstract and assigning FoR codes (Witten & Frate, 2005). The classifier could be made publically available allowing institutions to test their allocations prior to submission. The existing database of previous submissions would provide a substantial training set on which the classifier could be tuned and tested. The performance of the classifier could be quantified and, should institutions find egregious errors the classifier, could be adjusted in a process of continual improvement. Such a classifier would effectively encode a public standard of what should appear in a given FoR code and eliminate the personal variability and bias that is an inherent aspect of employing human panels to make these determinations.

It is essential that the ERA processes be strengthened. A lot is at stake for institutions' reputations when n/r outcomes occur, and this is especially unfair if the outcomes appear to be beyond their control or they are unable to rectify in subsequent processes. It is also essential in order to maintain the confidence of the sector that ERA processes are legitimate and functioning well. This is especially pertinent given the expanded remit to measure 'research engagement and impact', which is arguably more fraught and contentious than 'research excellence', in the forthcoming 2018 round.

Dr Paul Henman is Associate Professor of Sociology and Social Policy at the University of Queensland. He has recently led a five-year ARC Discovery project on performance measurement in Australia's health, school and university sectors.
Contact: p.benman@uq.edu.au

Professor Scott D. Brown is an ARC Future Fellow in the School of Psychology, University of Newcastle. His research interests include cognitive science, mathematical psychology, and neuroscience.

Professor Simon Dennis is Head of School of Psychology at the University of Newcastle, Australia. His areas of expertise include human memory and computational linguistics.

References

Adler, M. (2003). A Socio-Legal Approach to Administrative Justice. *Law & Policy, 25*(4), 323-352.

Australian Bureau of Statistics (ABS). (2008). *Australian and New Zealand Standard Research Classification*, Cat. No. 1297.0, Canberra: ABS.

Australian Research Council (ARC). (2014). *ERA 2015: Submission Guidelines.* Canberra: ARC.

Australian Research Council (ARC). (2015). *CEO Statement: ERA 2015 Speculation*. Retrieved from http://www.arc.gov.au/era-2015-speculation.

Bevan, G. & Hood, C. (2006). What's measured is what matters: targets and gaming in the English public health care system. *Public Administration, 84*(3), 517-538.

Bonnell, A.G. (2016). Tide or tsunami? The impact of metrics on scholarly research. *Australian Universities' Review 58*(1), 54-61.

Bovens, M., Goodin, R. E., & Schillemans, T. (eds.). (2014). *The Oxford handbook of public accountability.* Oxford: OUP.

Congdon, P. (2006). *Bayesian statistical modelling.* 2nd ed. Chichester: John Wiley & Sons.

Du Gay, P. (2000). *In praise of bureaucracy: Weber-organization-ethics.* London: Sage.

Elton, L. (2000). The UK research assessment exercise: unintended consequences. *Higher Education Quarterly, 54*(3), 274-283.

Gable, A. (2013). *ERA and the performance regime in Australian Higher Education: a review of the policy context*. Social Policy Unit Research Paper No. 6, Brisbane: University of Queensland.

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*(4), 277-304.

Haslam, N. & Koval, P. (2010). Possible research area bias in the Excellence in Research for Australia (ERA) draft journal rankings. *Australian Journal of Psychology, 62*(2), 112-114.

Hood, C. (2006). Gaming in targetworld: The targets approach to managing British public services. *Public Administration Review, 66*(4), 515-521.

Hunter, I. (1993) Bureaucrat, critic, citizen. *Arena Journal, 2*, 77-101.

Jacob, B.A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5), 761-796.

Knott, M. (2015). Warning over 'gaming' of research score. *The Sydney Morning Herald,* 18 November.

Kwok, J.T. (2013). *Impact of ERA research assessment on university behaviour and their staff.* Melbourne: NTEU National Policy and Research Unit.

Lathrop, D. & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice.* New York: O'Reilly Media, Inc.

Loussikian, K. (2015). Code flaws in ERA submissions. *The Australian*, 9 December.

Loussikian, K. (2016). 'Errors' in ERA subs behind rejections. *The Australian*. 16 March.

Martin, B.R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?. *Research Evaluation, 20*(3), 247-254.

Mashaw, J.L. (1985). *Bureaucratic justice.* Westford, MA: Yale University Press.

Vanclay, J.K. (2011). An evaluation of the Australian Research Council's journal ranking. *Journal of Informetrics 5*(2), 265-274.

Weber, M. (2009). *From Max Weber: Essays in sociology.* London: Routledge.

Witten, I.H. & Frate, E. (2005) *Data Mining: Practical machine learning tools and techniques.* 2nd ed. San Francisco, CA: Morgan Kaufmann.