Cogent Psychology

ISSN: 2331-1908 (Online) Journal homepage: www.tandfonline.com/journals/oaps20

Psychology and research assessment in the United Kingdom

Matthew Inglis, Colin Foster, Hugues Lortie-Forgues, Victoria Simms & Elizabeth Stokoe

To cite this article: Matthew Inglis, Colin Foster, Hugues Lortie-Forgues, Victoria Simms & Elizabeth Stokoe (2025) Psychology and research assessment in the United Kingdom, Cogent Psychology, 12:1, 2570100, DOI: 10.1080/23311908.2025.2570100

To link to this article: https://doi.org/10.1080/23311908.2025.2570100

Published online: 10 Oct 2025.

COGENT PSYCHOLOGY
2025, VOL. 12, NO. 1, 2570100
https://doi.org/10.1080/23311908.2025.2570100

SOCIAL PSYCHOLOGY | RESEARCH ARTICLE

OPEN ACCESS Check for updates

Psychology and research assessment in the United Kingdom

Matthew Inglis[a], Colin Foster[a], Hugues Lortie-Forgues[a], Victoria Simms[b] and Elizabeth Stokoe[c]

[a]Centre for Mathematical Cognition, Loughborough University, Loughborough, UK; [b]Department of Psychology, University of Ulster, Coleraine, UK; [c]Department of Psychological and Behavioural Science, The London School of Economics and Political Science, London, UK

ABSTRACT

What can we learn about psychology research in the UK, and its perceived quality, from examining manuscripts submitted to the psychology, psychiatry and neuroscience subpanel of the 2021 Research Excellence Framework (REF2021)? Using a latent Dirichlet allocation topic modelling approach, we identified 33 topics which collectively summarised the content of the journal articles returned to the subpanel. We found that the composition of submissions to the subpanel, in terms of these topics, explained a large proportion of the variance in the quality assessments they received from the expert peer review subpanel. Our model identified topics which were typically associated with receiving higher and lower unit-level quality assessments. In our discussion we pay particular attention to the fate of qualitative research, and discuss possible accounts for why units who returned a large amount of qualitative work tended to receive lower quality assessments than those who did not.

ARTICLE HISTORY
Received 1 June 2025
Revised 25 September 2025
Accepted 29 September 2025

KEYWORDS
Research; quality; assessment; research approaches; excellence

SUBJECTS
Medicine; Health and Social Care; Health & Society; Children and Youth; Medicine; Health and Social Care; Nursing; Learning Disabilities; Social Sciences; Behavioral Sciences; Developmental Psychology; Child Development; Social Sciences; Behavioral Sciences; Developmental Psychology; Cognitive Development

Research assessment in the UK

Since 1986, higher education institutions in the UK have been subject to evaluations of their research by the higher education funding councils (for a history, see Bence & Oppenheim, 2005). The 1986 'Research Selectivity Exercise' (RSE) evolved over six exercise periods into the 'Research Assessment Exercise' (RAE) in 2008 and the 'Research Excellence Framework' (REF) in 2014. The most recent exercise took place in 2021, with outcomes published in 2022. Because these assessments both inform research funding allocations and influence institutional reputations, they are taken remarkably seriously.

Above and beyond the global naming changes of the process, there have been substantive changes in rules and requirements for submissions (Marques et al., 2017). These have evolved from a relatively 'quick and dirty' (Jones & Sizer, 1990, p. 310) first evaluation in 1986 through to the introduction of complex regulations on the inclusion/exclusion of staff, the minimum/maximum numbers of outputs (the generic term for journal articles, books, chapters, conference proceedings, etc.), and the introduction of research environment statements. During this time, there have been substantial changes in the way disciplines – including psychology – are represented, categorized, and thus evaluated in these exercises. For instance, in the 1986 RAE, there were 36 'cost centres', in 1996 there were 60 panels, in REF 2014 there were 36 subpanels, and in REF 2021 there were 34 subpanels. In short, it is broadly recognised that the

CONTACT Matthew Inglis m.j.inglis@lboro.ac.uk Centre for Mathematical Cognition, Loughborough University, Loughborough, UK

research assessment process has had a 'pivotal role … in shaping academic disciplines in UK universities over the past decade' (Munoz-Chereau & Wyse, 2023, p. 7).

In the most recent REF 2021, groups of academics were submitted as 'units' to one of 34 assessment subpanels, defined either by disciplines or groups of related disciplines.[1] Each researcher was required to submit between 1 and 5 outputs, with each unit submitting an average of 2.5 outputs per researcher. Notably, psychology formed part of the psychology, psychiatry and neuroscience subpanel, to which 9773 outputs were returned, making it the fifth largest subpanel by output numbers. In contrast, many other subpanels served what many might view as single disciplines and were much smaller (e.g. archaeology [1208 outputs] and classics [1068 outputs]). In several cases, apparently related disciplines were placed in different subpanels (e.g. the geography and environmental studies subpanel [4479 outputs] and the earth systems and environmental sciences subpanel [4385 outputs]), sometimes following representations from learned societies (e.g. Royal Geographical Society, 2017).

Once submitted, these research outputs were assessed for their quality by a subpanel of senior academics and other experts. Each output was given a score for its quality in terms of 'originality, significance and rigour' on a five-point scale: ranging from unclassified to the highest 4*. These scores were combined to produce an output quality profile for each unit, which contributed to the overall quality profile for a unit, alongside analogous profiles for the reach and significance of the unit's impact (assessed via case studies) and the extent to which the unit's environment had been conducive to producing high-quality research (assessed via narrative environment statements and various metrics). One convenient way of expressing a unit's REF quality profile is to calculate a grade point average (GPA). The overall quality profile calculation in 2021 was weighted towards contributions from the assessment of outputs, making up 60% of the overall score, with impact contributing 25% and environment 15%. Therefore, within institutions substantial effort was spent on selecting the outputs to be included in unit submissions that were deemed most likely to receive high grades. Analyses have suggested that, in the REF 2014 assessment, a 'world leading' (4*) output 'earns between £7504 and £14,639 per year within the REF cycle' for an institution, with the caveat of between-discipline variability (Koya & Chowdhury, 2017).

Psychology and the REF

There have been several critiques of the positioning of psychology within a single subpanel alongside psychiatry and neuroscience, which many would view as separate disciplines. For instance, Langdridge (2020) described the psychology, psychiatry and neuroscience subpanel as 'a bit of an odd unit as most are single discipline' and suggested that this structure introduces bias into the assessment process: 'Personally, like many other psychologists, I think it's not the best situation as it leads to a problematic bias in favour of biological and cognitive psychology within our discipline'. One goal of the study reported in the current paper is to evaluate whether the results of the REF peer review process are consistent with this suspicion of bias.

If there are perceived biases *towards* certain topics, such as biological and cognitive psychology, then there may also be perceived biases *against* certain topics. Wetherell (2011) suggested that the REF structure has had a particularly negative effect on social psychology: 'The powerful narrative on which social psychology was once based is fragmenting in part due to Research Assessment Exercise (RAE/REF) pressures. Social psychological topics and research are migrating outside institutional Psychology' (p. 399). Wetherell's suggestion was that many social psychologists' work was submitted to other subpanels, such as sociology, or communication, cultural and media studies. In a similar vein, Collins and Bunn (2016) argued that, with the advent of the REF, work on the history of psychology has been marginalised.

Some existing data supports Wetherell's (2011) suggestion that many psychologists have been returned to subpanels other than psychology, psychiatry and neuroscience. In response to REF2014, research undertaken by the Research Board of the British Psychological Society (BPS) indicated that 78.5% of (self-identified) psychology researchers who responded to their survey were submitted to the psychology, psychiatry and neuroscience subpanel, with the remaining respondents being submitted to six different subpanels (Research Board of the British Psychological Society, 2014). Some respondents – 10.4% – were not happy with the choice of subpanel that they were submitted to, with some suggesting that psychology should have its own subpanel (Research Board of the British Psychological Society, 2014). The

BPS Research Board report concluded that qualitative and social psychology researchers were less likely to be submitted to the REF, which may indicate 'a worrying trend in terms of the sub-field and methodological bias in returns to the REF' (p. 8).

The BPS Social Psychology Section ran a separate survey and largely replicated these findings, indicating that social psychologists who conducted qualitative research were less likely to be submitted to the REF than their quantitative or mixed-methods peers (Research Board of the British Psychological Society, 2014). Respondents also indicated that they believed that decision making may have been based on risk aversion of REF submission leaders, due to perceived biases in favour of quantitative methods within the psychology, psychiatry and neuroscience panel. In response to these concerns, the BPS Qualitative Methods in Psychology section produced 'a pragmatic tool to support qualitative psychologists in the United Kingdom who are obliged to produce outputs for submission to REF' (Brooks et al., 2018, p. 4).

In sum, there seems to be substantial concern that psychology's inclusion in a subpanel alongside psychiatry and neuroscience has led to parts of the discipline being disadvantaged in myriad ways.[2] However, official statements by those involved in the process have sought to reassure the community that this is not the case. For instance, in advance of REF2021, Susan Gathercole, chair of the psychology, psychiatry and neuroscience subpanel, publicly stated that 'the panel recognise qualitative research as a key approach in many areas of psychology and one in which many UK researchers excel' (Brooks et al, 2018, p. 5). In addition, following the publication of the REF2021 results, the subpanel highlighted in their report that research excellence was observed across outputs that took a variety of approaches, including those which used advanced quantitative and qualitative research methodologies, experimental designs, those which worked with clinical populations and those which had clear pathways to societal impact (REF, 2022).

What this discussion lacks is data on the composition of submissions to the subpanel, and the extent to which submissions that focused on different topics received different quality profiles. Our goal in the current study was to conduct an analysis of these issues. We addressed three main questions. First, we sought to quantitatively analyse the content of submissions made to the REF2021 psychology, psychiatry and neuroscience subpanel, particularly in terms of their outputs' substantive focus and methodological choices. Second, we assessed the extent to which units' quality profiles could be predicted by the focus and methodological makeup of their outputs. Third, we explored whether there had been changes between REF2014 and REF2021 in terms of either the content of outputs submitted to the psychology, psychiatry and neuroscience subpanel, or the panels' approaches to peer review.

The current study

We adopted a similar method to that used by Inglis, Foster, Lortie-Forgues and Stokoe (2024) in their analysis of returns to the REF2021 education subpanel. Inglis et al. used a machine learning approach known as latent Dirichlet allocation topic modelling to identify the main topics written about in journal articles returned to the education subpanel. From this they were able to identify those topics which had positive associations with the REF peer review outcomes at the unit level. In other words, they found that some topics (notably large-scale secondary data analyses) were typically returned in larger proportions by units that received high scores from the REF panel, and that other topics (notably interview studies) were typically returned in larger proportions by units that did less well. If concerns about possible biases of the psychology, psychiatry and neuroscience panel discussed above have any merits, we might expect to see analogous associations, consistent with the hypothesised biases.

Topic modelling seeks to identify the themes, or topics, contained within a large collection of texts (Blei, Ng, & Jordan, 2003). For example, if a document contains many instances of the words 'keyboard', 'screen' and 'mouse', we might infer that the document is, to some extent at least, about computers. Formally, a topic is defined to be a probability distribution over words. So, a computer topic would associate high probabilities with words related to computers ('keyboard', 'screen', 'mouse'), and low probabilities with unrelated words ('toothpaste', 'coronation', 'frost').

Considering the process in reverse helps to elucidate the method. Imagine that we have a set of topics and wish to produce documents. If we wanted to write a document that is 70% about computers, 20% about Greece and 10% about emotions, then whenever we added a word to our document we

would select it from the computers topic with probability 0.7, from the Greece topic with probability 0.2, and from the emotions topic with probability 0.1. The topics are themselves probability distributions over words: perhaps the emotions topic assigns the word 'unhappy' a probability of 0.002. If so, then each time we added a word to our document the probability of it being 'unhappy' (from the emotions topic) would be 0.1×0.002. Two considerable simplifications are made when constructing documents in this manner: both word order and so-called 'stop words' – words that do not convey semantic content, such as 'the', 'as' and 'is' – are ignored.

Topic modelling assumes that a specified set of documents was created using this method and then attempts to identify the most plausible topics. This allows the composition of each document to be specified. For instance, we might conclude that a document is made up of 25% of words from topic 1, 10% of words from topic 2 and so on (these percentages represent the number of words from each topic after the removal of stop words).

Method

In total, 9773 outputs were submitted to the psychology, psychiatry and neuroscience subpanel REF2021. Of these, 9753 (99.8%) were declared to be journal articles by the submitting units, and all but one of them were written in English. We were able to obtain pdf versions of 9691 (99.4%) of these, which were converted to plain text using the UNIX pdftotext command (Poppler, 2022). We then used MALLET (version 2.0.8RC2, McCallum, 2002) to calculate possible topic models, applying MALLET's default list of stop words.

To assess how many topics to use in our primary analysis, we adopted a perplexity method (Blei, Ng, & Jordan, 2003; Jacobi, Van Atteveldt, & Welbers, 2018). The corpus was split into a training section (80%) and a testing section (20%), and topic models with 10 topics, 20 topics, … 100 topics were fitted to the training section. Perplexity, an estimate of model fit (with lower values indicating better fit), was then calculated using the testing section. Figure 1 shows perplexities for each model. Jacobi et al. suggested basing a choice of topic number on where this graph 'levels off', in a similar manner to a scree plot in an exploratory factor analysis. Given the piecewise linear regression shown in Figure 1, we opted for a model with 33 topics for our primary analysis.

Our 33-topic model provided us with the topic-by-topic composition of each of 9691 English-language journal articles returned to the subpanel. To illustrate, consider Simms et al.'s (2015) article 'Nature and origins of mathematics difficulties in very preterm children: a different etiology than developmental dyscalculia'. The article characterised the difficulties preterm children have when learning mathematics in school, and demonstrated that these were different from those faced by children with developmental dyscalculia. Our model identified that 65.1% of the article's words came from Topic 27 and 19.9% from Topic 9 (here, and throughout the rest of the paper, the percentages of a paper's words from a given topic are given after the removal of stop words). Using the process described below, these topics were named 'Development, Lifespan and Developmental Differences' and 'Child Psychology and Psychiatry', respectively, which seems to appropriately capture the content of Simms et al.'s article.

Next, we assessed the model by calculating each output's most characteristic topic (the topic from which it drew the highest proportion of words). The mean percentage of words from the most

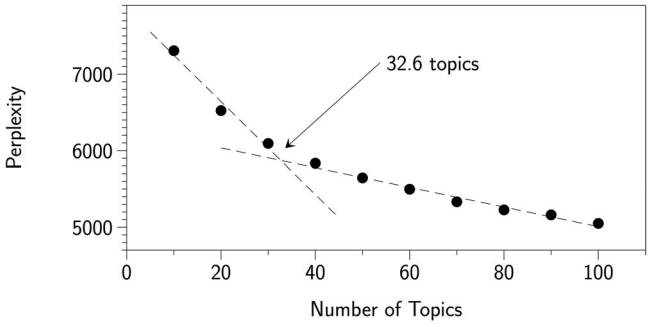

Figure 1. Perplexities associated with models with 10, 20, 30, …, 100 topics. The dotted lines show a one-break piecewise linear regression line of best fit.

characteristic topic was 51.5% (SD 15.9%), and the mean percentage of words from the second and third most characteristic topics were 21.4% (SD 8.6%) and 11.0% (SD 5.5%) respectively. In other words, it was typical for most outputs to be well characterised by a small number of topics.

Results

Table 1 shows the characteristic words for each of the 33 topics, alongside the article that had the highest proportion of words from the topic, the name we gave to describe the topic, and the topic's mean proportion of words (averaged over all papers). Names were assigned based on the characteristic words and, where that was insufficient, a careful reading of papers with particularly high proportions of words from the topic. In all cases it was relatively straightforward to assign names to topics. The topic compositions for each of the 9691 journal articles is available in the associated online materials at http://doi.org/10.17028/rd.lboro.28881821. Studying these data, in conjunction with Table 1, will permit readers to assess whether they feel our topic names appropriately capture the meaning of each topic.[3]

Next, we calculated each submitted unit's mean proportion for each topic. This gave us a measure of the character of each unit's submission. For instance, 13.8% of the University of Kent's 'composite mean paper' (an imagined paper composed of the same topic weightings as the mean topic weightings of the actual papers returned by the University of Kent) was made up of words from the Experimental and Evolutionary Social Psychology topic (Kent's most prevalent topic). Similarly, the most prevalent topics from the University of Ulster's and University of York's composite mean papers were respectively the Mental Health Epidemiology topic (21.9%) and the Face Recognition topic (14.5%). These results seem consistent with our impressions of these departments' research strengths, providing some support for the face validity of our model. We identified the most prevalent topic from each unit's composite mean paper. The mean percentages of words from the most prevalent topics was 15.7% (SD 7.0%), and the mean percentages of words from units' second and third most prevalent topics were 10.6% (SD 3.1%) and 8.6% (SD 1.9%) respectively. The mean topic weightings, across all topics, for each institution submitted to the psychology, psychiatry and neuroscience subpanel, alongside their output quality profiles, output GPA, and number of FTE staff submitted, are available in the associated online materials.

The overall mean proportions of words from each topic, across all of the papers we analysed, are shown in Table 1. These figures give an overall sense of the balance of topics represented in articles submitted to the psychology, psychiatry and neuroscience subpanel in REF2021. Statistical and Mathematical Modelling was the most prevalent topic (6.4%), followed by Qualitative Research (4.8%), and Neuroimaging (4.8%). Of the non-methods topics, the most prevalent were Personality and Individual Differences (4.1%), Neurological Conditions (4.1%) and Visual Cognition and Attention (4.0%).

Next, we evaluated the extent to which the topic proportions of the composite mean paper submitted by each unit could account for the unit-level output GPAs assigned by the subpanel. Because topic proportions sum to 1, analysing these data using a standard regression approach would be impossible due to perfect multicollinearity. Instead, we adopted the compositional base 2 additive log-ratio regression approach advocated by Coenders and Pawlowsky-Glahn (2020). Since we analysed 99.4% of the journal articles returned to the subpanel, we conceptualise this regression as being a whole-population analysis (Berk, 2004, p. 42) and therefore do not report inferential statistics. Given this, although our analysis permits conclusions to be drawn about REF2021, it does not allow us to assess whether the resulting model can accurately predict judgements of research quality made outside this context. Later in the paper we return to this issue by using our model to analyse REF2014 submissions.

This regression provided two main results. First, we assessed the overall model fit, which indicates how much of the variance in output GPAs can be collectively explained by the 33 topics. We also ran a model in which the proportion of 4* outputs was the dependent variable, which yielded very similar results.[4] Our model explained a very large proportion of the variance in output GPAs, $R^2 = 0.901$. However, given the low number of units (93) relative to the number of topics needed to characterise research returned to the subpanel (33), it is possible that this very large R^2 is the result of overfitting. To assess this possibility, we ran a Leave-One-Out Cross-Validation (LOOCV) analysis. We ran 93 separate regressions, in which each unit was excluded in turn. The regression coefficients from these were used to predict the excluded unit's output GPA. Once we had calculated a predicted output GPA for each unit,

Table 1. The 33 topics in our model, together with their characteristic words, the mean percentage of words from each topic (averaged over outputs) in REF2021 and REF2014, and the paper with the highest proportion of words from the topic in REF2021.

Topic	Name	Words	REF2021 Mean %	REF2014 Mean %	Paper with highest proportion of words from the topic
1	Law and human behaviour	Participants condition moral study information social psychology research effect people journal person e.g. game questions conditions behavior interview studies effects	2.411	1.827	Tekin, S., Granhag, P. A., Strömwall, L., Giolla, E. M., Vrij, A., & Hartwig, M. (2015). Interviewing strategically to elicit admissions from guilty suspects. *Law and Human Behavior, 39*(3), 244–252.
2	EEG and stimulation	Stimulation activity brain EEG power cortex time amplitude neural frequency analysis data effects response trials visual left alpha ERP effect	2.793	2.633	Parkin, B. L., Bhandari, M., Glen, J. C., & Walsh, V. (2019). The physiological effects of transcranial electrical stimulation do not apply to parameters commonly used in studies of cognitive neuromodulation. *Neuropsychologia, 128,* 332–339.
3	Health, exercise and appetite	Pain food body alcohol participants eating weight study consumption BMI activity drinking intake obesity control effects physical health exercise effect	2.188	1.816	Dalton, M., Hollingworth, S., Blundell, J., & Finlayson, G. (2015). Weak satiety responsiveness is a reliable trait associated with hedonic risk factors for overeating among women. *Nutrients, 7*(9), 7421–7436.
4	Mitochondrial disorders	Cells cell expression fig protein genes control proteins figure human gene levels analysis data mice mouse RNA neurons mutant mitochondrial	3.848	3.547	Ivankovic, D., Chau, K. Y., Schapira, A. H., & Gegg, M. E. (2016). Mitochondrial and lysosomal biogenesis are activated following PINK 1/parkin-mediated mitophagy. *Journal of Neurochemistry, 136*(2), 388–402.
5	Negative emotions and anxiety	Anxiety emotional negative stress emotion participants positive depression cognitive study psychological research journal fear scale control PTSD mindfulness group affective	3.041	2.693	Bailey, R., & Wells, A. (2016). Is metacognition a causal moderator of the relationship between catastrophic misinterpretation and health anxiety? A prospective study. *Behaviour Research and Therapy, 78*, 43–50.
6	Stem cell therapies	Cells cell fig retinal mice human mutations prion eye protein mouse patients muscle figure gene retina data control performed loss	1.835	1.374	Gonzalez-Cordero, A., Kruczek, K., Naeem, A., Fernando, M., Kloc, M., Ribeiro, J., … & Ali, R. R. (2017). Recapitulation of human retinal development from human pluripotent stem cells generates transplantable populations of cone photoreceptors. *Stem Cell Reports, 9*(3), 820–837.
7	Mental health epidemiology	Health mental risk study data suicide age smoking years research population people factors violence table mortality cannabis prevalence alcohol national	3.778	3.064	Clements, C., Hawton, K., Geulayov, G., Waters, K., Ness, J., Rehman, M., … & Kapur, N. (2019). Self-harm in midlife: analysis using data from the Multicentre Study of Self-harm in England. *British Journal of Psychiatry, 215*(4), 600–607.
8	Visual cognition and attention	Task trials target participants experiment attention effect response visual stimuli http://dx.doi.org control effects condition performance tasks trial time attentional stimulus	4.019	6.072	Grubert, A., & Eimer, M. (2015). Rapid parallel attentional target selection in single-color and multiple-color visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(1), 86–101.
9	Child psychology and psychiatry	Child children age childhood study ADHD parents years adolescents problems maternal development early adolescent symptoms family school risk psychiatry sample	3.518	3.158	Grabow, A. P., Khurana, A., Natsuaki, M. N., Neiderhiser, J. M., Harold, G. T., Shaw, D. S., … & Leve, L. D. (2017). Using an adoption–biological family design to examine associations between maternal trauma, maternal depressive symptoms, and child internalizing and externalizing behaviors. *Development and Psychopathology, 29*(5), 1707–1720.

(*Continued*)

Table 1. Continued.

Topic	Name	Words	REF2021 Mean %	REF2014 Mean %	Paper with highest proportion of words from the topic
10	Action and motor perception	Motor action movement participants hand movements body actions visual control touch perception tactile condition task imagery virtual doi spatial conditions	2.353	3.017	Perera, A. T. M., Newport, R., & McKenzie, K. J. (2017). Changing hands: persistent alterations to body image following brief exposure to multisensory distortions. *Experimental Brain Research, 235*, 1809–1821.
11	Chemical influences on cognition	Sleep mice rats animals day group time circadian test min learning light insomnia memory effect effects hippocampal conditioning behavioral groups	1.485	1.736	Pilorz, V., Tam, S. K., Hughes, S., Pothecary, C. A., Jagannath, A., Hankins, M. W., … & Peirson, S. N. (2016). Melanopsin regulates both sleep-promoting and arousal-promoting responses to light. *PLOS Biology, 14*(6), e1002482.
12	Systematic reviews and meta-analyses of health interventions	Studies risk review data bias study interventions patients cancer health intervention quality outcomes low care effect meta-analysis included treatment reported	2.194	1.503	Davey, P., Brown, E., Charani, E., Fenelon, L., Gould, I. M., Holmes, A., … & Wilcox, M. (2013). Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database of Systematic Reviews, https://doi.org/10.1002/14651858.CD003543.pub4
13	Psycholinguistics	Language word words reading semantic speech effects processing lexical e.g. effect phonological sentence English comprehension participants frequency sentences learning model	2.546	3.591	Cai, Z. G., Pickering, M. J., Wang, R., & Branigan, H. P. (2015). It is there whether you hear it or not: Syntactic representation of missing arguments. *Cognition, 136*, 255–267.
14	Memory	Memory recall participants retrieval items task performance working learning encoding effect memories recognition experiment test episodic study effects item presented	2.510	3.436	Cortis Mack, C., Dent, K., & Ward, G. (2018). Near-independent capacities and highly constrained output orders in the simultaneous free recall of auditory-verbal and visuo-spatial stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(1), 107.
15	Genome-wide studies	Genetic university genes gene association data variants research department analysis supplementary study institute risk genetics genome-wide USA loci psychiatry SNPS	3.143	2.197	Pantelis, C., Papadimitriou, G. N., Papiol, S., Parkhomenko, E., Pato, M. T., Paunio, T., … & O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421–427.
16	Neurological conditions	Patients disease dementia clinical study brain stroke Alzheimer's cognitive age Parkinson's neurology years neurol controls patient impairment group epilepsy data	4.104	4.482	Jabbari, E., Holland, N., Chelban, V., Jones, P. S., Lamb, R., Rawlinson, C., … & Morris, H. R. (2020). Diagnosis across the spectrum of progressive supranuclear palsy and corticobasal syndrome. *JAMA Neurology*, *77*(3), 377–387.
17	Thinking and reasoning	Participants psychology belief cognitive experiment causal people events theory evidence condition reasoning e.g. thinking implicit psychological cognition http://dx.doi.org information beliefs	1.819	2.255	Johnson, S. G., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology, 89*, 39–70.
18	Theraputic interventions	Treatment intervention trial health group study months therapy participants care baseline outcome data research follow-up trials primary outcomes score analysis	3.592	2.656	Everitt, H. A., Landau, S., O'Reilly, G., Sibelli, A., Hughes, S., Windgassen, S., … & Moss-Morris, R. (2019). Cognitive behavioural therapy for irritable bowel syndrome: 24-month follow-up of participants in the ACTIB randomised trial. *The Lancet Gastroenterology & Hepatology, 4*(11), 863–872.

(*Continued*)

Table 1. Continued.

Topic	Name	Words	REF2021 Mean %	REF2014 Mean %	Paper with highest proportion of words from the topic
19	Central nervous system conditions	cells mice fig cell spinal neurons cord expression brain control microglia injury nerve animals mouse axons astrocytes tissue data pain	2.186	2.023	Bartus, K., James, N. D., Didangelos, A., Bosch, K. D., Verhaagen, J., Yánez-Munoz, R. J., … & Bradbury, E. J. (2014). Large-scale chondroitin sulfate proteoglycan digestion with chondroitinase gene therapy leads to reduced pathology and modulates macrophage phenotype following spinal cord contusion injury. *Journal of Neuroscience, 34*(14), 4822–4836.
20	Face recognition	Face faces facial recognition participants images stimuli processing expressions emotion image perception social expression experiment effect test identity familiar psychology	2.349	2.464	Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, *82*, 48–62.
21	Early development including comparative cognition	Children infants development social child learning children's age condition infant object study developmental months doi test gestures human task communication	2.519	2.376	Roberts, A. I., Vick, S. J., Roberts, S. G. B., & Menzel, C. R. (2014). Chimpanzees modify intentional gestures to coordinate a search for hidden food. *Nature Communications*, *5*(1), 3088.
22	Personality and individual differences	Personality study social items model psychology journal research factor http://dx.doi.org scale psychological participants effects measures behavior positive values e.g. relationship	4.112	3.518	Buchanan, K., & Bardi, A. (2015). The roles of values, behavior, and value-behavior fit in the relation of agency and communion to well-being. *Journal of Personality*, *83*(3), 320–333.
23	Experimental social and evolutionary psychology	Social group women groups identity study contact sexual men sex intergroup individuals gender psychology political participants collective doi attitudes effect	2.628	2.320	Cakal, H., Halabi, S., Cazan, A. M., & Eller, A. (2021). Intergroup contact and endorsement of social change motivations: The mediating role of intergroup trust, perspective-taking, and intergroup anxiety among three advantaged groups in Northern Cyprus, Romania, and Israel. *Group Processes & Intergroup Relations*, *24*(1), 48–67.
24	Brain structure	Figure figures cortical http://dx.doi.org brain cite n.s. press human m.a j.m development m.j a.m age j.a current biology supplemental Elsevier	1.189	0.988	Lunnon, K., Keohane, A., Pidsley, R., Newhouse, S., Riddoch-Contreras, J., Thubron, E. B., … & AddNeuroMed Consortium. (2017). Mitochondrial genes are altered in blood early in Alzheimer's disease. *Neurobiology of Aging*, *53*, 36–47.
25	Cellular neuroscience	Neurons cells figure activity synaptic mice neuron cell firing fig cortex doi neurosci neuronal recordings interneurons plasticity spike stimulation responses	3.064	3.018	Tigaret, C. M., Olivo, V., Sadowski, J. H., Ashby, M. C., & Mellor, J. R. (2016). Coordinated activation of distinct Ca2+ sources and metabotropic glutamate receptors encodes Hebbian synaptic plasticity. *Nature Communications*, *7*(1), 10289.
26	Neuroimaging	Brain cortex regions functional connectivity left FMRI gyrus temporal network analysis frontal imaging activation cortical neuroimage neural areas matter anterior	4.769	5.913	Jackson, R. L., Hoffman, P., Pobric, G., & Ralph, M. A. L. (2016). The semantic network at work and rest: differential connectivity of anterior temporal lobe subregions. *Journal of Neuroscience*, *36*(5), 1490–1501.
27	Development, Lifespan and Developmental Differences	children age autism group cognitive adults asd performance older developmental groups ability study scores differences measures development test years task	3.612	3.931	McPhillips, M., Finlay, J., Bejerot, S., & Hanley, M. (2014). Motor deficits in children with autism spectrum disorder: A cross-syndrome study. *Autism Research*, *7*(6), 664-676.

(*Continued*)

Table 1. Continued.

Topic	Name	Words	REF2021 Mean %	REF2014 Mean %	Paper with highest proportion of words from the topic
28	Statistical and mathematical modelling	model data models fig number analysis time information values set methods results figure distribution size parameters individual https://doi.org average based	6.366	5.405	Wang, J., Tian, F., Yu, H., Liu, C. H., Zhan, K., & Wang, X. (2017). Diverse non-negative matrix factorization for multiview data representation. *IEEE Transactions on Cybernetics*, *48*(9), 2620–2632.
29	Qualitative research	Research social people health participants work support experience analysis experiences information qualitative university change journal online psychology data time process	4.834	4.037	Oakley, L., Fenge, L. A., & Taylor, B. (2022). 'I call it the hero complex'–Critical considerations of power and privilege and seeking to be an agent of change in qualitative researchers' experiences. *Qualitative Research in Psychology*, *19*(3), 587–610.
30	Psychopharmacology	Effects levels blood treatment effect cortisol receptor brain drug response study min dopamine placebo activity dose oxytocin concentrations stress increased	2.269	3.021	Yue, J. T., Abraham, M. A., LaPierre, M. P., Mighiu, P. I., Light, P. E., Filippi, B. M., & Lam, T. K. (2015). A fatty acid-dependent hypothalamic–DVC neurocircuitry that regulates hepatic secretion of triglyceride-rich lipoproteins. *Nature Communications*, *6*(1), 5970.
31	Psychiatric Disorders	Depression disorder psychiatry patients schizophrenia symptoms disorders clinical psychosis study depressive psychiatric psychotic group treatment bipolar studies participants risk individuals	3.469	3.583	Lin, A., Wood, S. J., Nelson, B., Beavan, A., McGorry, P., & Yung, A. R. (2015). Outcomes of nontransitioned cases in a sample at ultra-high risk for psychosis. *American Journal of Psychiatry*, *172*(3), 249–258.
32	Low-level cognition processes	Learning reward trials choice participants decision task trial model outcome figure time response choices prediction error fig experiment cue effect	2.439	2.033	Buckley, M. G., Smith, A. D., & Haselgrove, M. (2016). Thinking outside of the box: Transfer of shape-based reorientation across the boundary of an arena. *Cognitive Psychology*, *87*, 53–87.
33	Sensory processing and perception	Visual auditory stimuli stimulus perception motion responses response noise speech perceptual spatial conditions vision presented sensory experiment sound temporal fig	3.018	4.315	Rocchi, F., Ledgeway, T., & Webb, B. S. (2018). Criterion-free measurement of motion transparency perception at different speeds. *Journal of Vision*, *18*(4), 5–5.

we correlated these with the actual output GPAs, as shown in Figure 2, finding a strong correlation, $r = 0.839$, $r^2 = 0.704$. In other words, our model was successful at predicting out-of-sample output GPAs, suggesting that the extremely high model fit was not due to overfitting. As discussed later in the paper, we also explored the out-of-sample predictiveness of our model by applying it to REF2014 and evaluating the extent to which it could predict judgements about different outputs made by a different panel.

The second main output from our compositional regression analysis consisted of the regression coefficients associated with each of the 33 topics, as shown in Table 2. In a base 2 additive log-ratio compositional regression, regression coefficients capture the expected change in the dependent variable if the value of the ratios between the given predictor and all other predictors doubles, with the other predictors retaining identical relative ratios. For instance, the regression coefficient associated with the Neuroimaging topic was 0.029. This means that if one unit's composite mean paper had twice as much content about neuroimaging (relative to the other topics) as another units, and if both had an identical balance across the other topics, we would predict that the first unit's output GPA would be 0.029 higher than the second's. The regression coefficients varied from 0.047 (Sensory Processing and Perception) to −0.092 (Qualitative Research).

Two topics had coefficients greater than 0.04: Sensory Processing and Perception, and Law and Human Behaviour. Papers with a particularly high proportion of words from the Sensory Processing and Perception topic typically focused on the low-level processes involved in vision. For instance, Rocchi, Ledgeway and Webb's (2018) investigation of motion transparency perception, published in the *Journal of Vision*, had 86% of its words from this topic. Characteristic papers from the Law and Human Behaviour topic

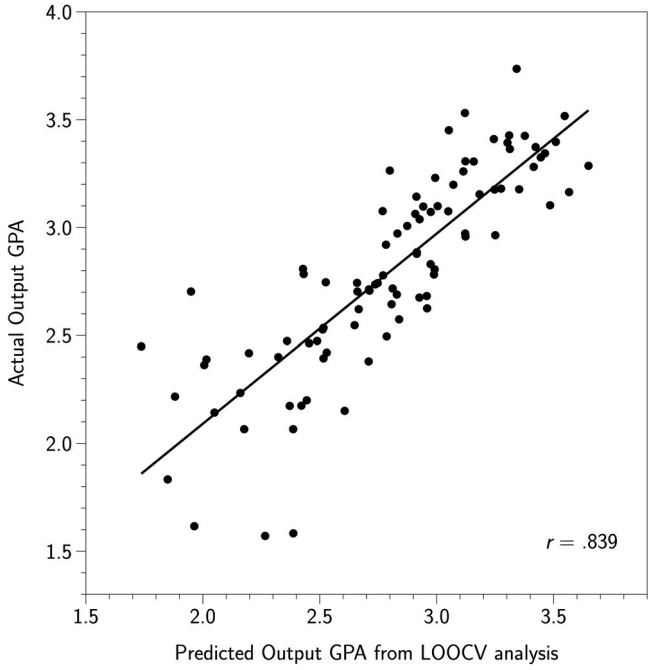

Figure 2. A plot showing units' actual output GPAs from REF2021 against the output GPAs predicted by our leave one out cross validation analysis.

typically reported experimental studies focused on legal/criminal issues. For instance, some papers focused on methods to elicit admissions from guilty suspects, and others on lie detection in various contexts. Papers with high proportions from this topic were often published in the journals *Law and Human Behaviour* or *Legal and Criminological Psychology*.

Three topics had regression coefficients below −0.04: Qualitative Research, Face Recognition, and Mental Health Epidemiology. Most notable was the Qualitative Research topic, which had the largest negative coefficient, −0.092. The articles with high proportions of words from this topic used qualitative methods to analyse various psychological issues. For instance, Lewis et al.'s (2017) article entitled 'Public sector austerity cuts in Britain and the changing discourse of work–life balance' reported insights from a series of interviews with senior human resources professionals and had 94% of its words from this topic. Many of the articles with high proportions of words from the qualitative topic were published in *Qualitative Research in Psychology*. The Face Recognition topic was characterised by words such as 'face', 'faces', 'facial' and 'recognition', and the paper with the highest proportion of words from the topic was Bobak et al.'s (2016) study of individuals with superior face recognition skills. The Mental Health Epidemiology topic captured research which investigated mental health issues using large-scale secondary data. The paper with the highest proportion of words from the topic was Clements et al.'s (2019) article 'Self-harm in midlife: analysis using data from the Multicentre Study of Self-harm in England'. Other papers with particularly high proportions of words from this topic analysed data from datasets generated by the Multicentre Study of Self-Harm, the Global Burden of Disease Study 2013, the Danish Civil Registration System, and various other large-scale cohort studies. A full dataset showing the topic proportions for each of the 9691 articles we analysed is available in the online materials. This dataset can be used to interrogate the accuracy of our topic name choices.

In sum, our regression explained a large proportion of the variance in units' output GPAs, including when we conducted a LOOCV analysis. This allowed us to identify those topics, methods and approaches that were, at the unit level at least, associated with judgements of higher and lower quality made by the REF2021 psychology, psychiatry and neuroscience subpanel. We found several topics that were associated with higher scores, and several, particularly Qualitative Research, that were associated with lower scores.

To address our remaining two research questions, concerning (i) changes to the focus of research returned to the psychology, psychiatry and neuroscience subpanel over time, and (ii) the extent to which

Table 2. A compositional regression predicting REF2021 output GPAs with our 33 topics.

Predictor	Regression coefficient
(Intercept)	3.176
Sensory processing and perception	0.047
Law and human behaviour	0.046
Genome-wide studies	0.036
Low-level cognition processes	0.034
Neuroimaging	0.029
Child psychology and psychiatry	0.027
Experimental social and evolutionary psychology	0.026
Development, lifespan and developmental differences	0.025
Mitochondrial disorders	0.024
Psychiatric disorders	0.022
Visual cognition and attention	0.022
Health, exercise and appetite	0.015
Chemical influences on cognition	0.013
Therapeutic interventions	0.013
Stem cell therapies	0.011
Brain structure	0.003
Cellular neuroscience	0.002
Psycholinguistics	0.000
Thinking and reasoning	−0.002
Action and motor perception	−0.003
Memory	−0.006
Central nervous system conditions	−0.007
EEG and stimulation	−0.007
Early development including comparative cognition	−0.009
Statistical and mathematical modelling	−0.014
Neurological conditions	−0.017
Personality and individual differences	−0.021
Systematic reviews and meta-analyses of health interventions	−0.028
Psychopharmacology	−0.033
Negative emotions and anxiety	−0.036
Mental health epidemiology	−0.055
Face recognition	−0.064
Qualitative research	−0.092
	$R^2 = 0.901$

Topics are ordered by the size of the regression coefficient.

the model can estimate out-of-sample peer review judgements, we applied our model to journal articles submitted to REF2014.

Applying the model to REF2014

Compared to 2021, fewer universities made returns to the psychology, psychiatry and neuroscience sub-panel in 2014 (82 compared to 93), and these returns contained slightly fewer outputs (9126 compared to 9773). Of these, 9086 (99.6%) were self-declared to be journal articles, all of which were written in English. We were able to obtain pdf copies of 8843 (97.3%). As before, we converted these 8843 articles into plain text using the UNIX pdftotext command (Poppler, 2022), and used our 33-topic REF2021 model to calculate the composition of each article. A full dataset showing the topic compositions for the 8843 articles in our REF2014 sample is available in the online materials.

We address two main questions. First, have there been changes in the prevalence of topics between the two REF exercises? Second, can our model successfully predict unit-level output GPAs achieved by the 2014 papers, as assigned by the REF2014 subpanel?

Recall that Wetherell (2011) had suggested that the RAE had led to social psychology research 'migrating' away from psychology submissions. Can we find evidence for the continuation of this suggested trend in our data? The mean proportion of words, averaged across all articles, from each topic in REF2014, is shown in the fifth column of Table 1. Some notable changes between 2014 and 2021 can be observed. Several traditional cognitive psychology topics have shown substantial declines in prevalence. For instance, the Visual Cognition and Attention, Sensory Processing and Perception, Psycholinguistics, and Memory topics all declined by over 25% between 2014 and 2021. In contrast, the Law and Human Behaviour, Stem Cell Therapies, Therapeutic Interventions, Genome-Wide Studies, and Systematic Reviews

and Meta-Analyses of Health Interventions topics all increased by over 30%. It seems that returns to the psychology, psychiatry and neuroscience unit of assessment have become less cognitive, and more applied in character over time. Notably, consistent with Wetherell's suggestion, the Experimental Social and Evolutionary Psychology topic also showed a decline in prevalence, from 2.6% to 2.3% (13%), although this was less pronounced than in some other cases.

Next, we calculated the mean composite paper associated with each of 81 of the 82 submissions made to the REF2014 subpanel in a similar manner to our REF2021 analysis. We excluded the 82nd submission, from Newman University, as it contained only 3.0 FTE staff and therefore an output quality profile was not published as part of the official REF data. We then used the regression coefficients for the REF2021 model shown in Table 2 to calculate predicted output GPAs. This gave estimates of the 2014 output GPAs that we would expect each submission to receive, based solely on our topic model and the associated regression coefficients from 2021. Next, we compared these predicted output GPAs with the actual output GPAs assigned by the REF2014 subpanel, as shown in Figure 3. The correlation between the predicted and actual output GPAs was high, at $r = 0.794$, $r^2 = 63.1\%$.

Importantly, our 2021 model explained more of the variance in 2014 output scores than is typically achieved by citation analyses. Pride and Knoth (2018) found a correlation of $r = 0.659$, $r^2 = 43.4\%$, between the median number of citations achieved by units' submitted papers as of 2014 and their output GPAs in the psychology, psychiatry and neuroscience subpanel. Our topic model explained around 50% more variance in REF2014 output GPAs than Pride and Knoth's citation methods. In addition, it might be possible to increase our r^2 further if we adopted a non-linear model, albeit at the cost of reducing interpretability.

In sum, we were able to produce an accurate estimate of how the units which made returns to REF2014 were assessed by the 2014 subpanel, confirming the results of our LOOCV analysis, which showed that the large r^2 observed for our REF2021 model was not simply due to overfitting. The fact that our REF2021 model was able to predict the outcomes of the REF2014 review process suggests that there was a reasonable degree of consistency in the approaches used by the two subpanels to assess research quality.

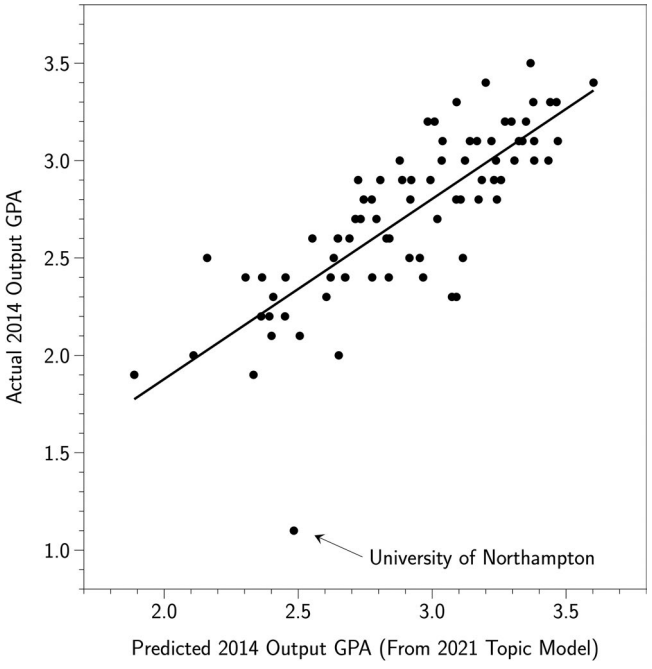

Figure 3. A plot showing units' actual output GPAs from REF2014 against the output GPAs predicted by our topic model.

Discussion

Summary of main findings

We explored the peer review process used by the psychology, psychiatry and neuroscience subpanel of REF2021. By studying the words used by the journal articles submitted to the panel, we were able to identify 33 main topics that collectively characterised the composition of outputs returned to the subpanel. These topics collectively explained a large proportion of the variance – 90.1% in our main analysis, 70.4% in our LOOCV analysis – in the subpanel's judgements of unit-level research quality. We found that units which submitted more work focused on Sensory Processing and Perception, and Law and Human Behaviour tended to receive higher output scores, and those which submitted more work focused on Mental Health Epidemiology, Face Recognition and, particularly, Qualitative Research, tended to receive lower output scores. Obviously, these relationships are correlational, and we cannot draw conclusions about causality. Nevertheless, given the speculation that including psychology in a subpanel alongside psychiatry and neuroscience has the effect of favouring some parts of the discipline over others, these relatively strong relationships require unpacking.

In the remainder of the paper, we discuss two main issues. First, we highlight that, by necessity, our analyses were conducted at the unit-level, and that we would expect to be able to account for a lower proportion of the variance of peer review scores if we had access to output-level data. Second, we consider the extent to which our findings support the hypothesis that the current structure of the psychology, psychiatry and neuroscience panel is biased against certain types of psychology research.

Ecological correlations

One limitation of our analysis is that, by necessity, we were forced to conduct analyses at the unit-level, not the output level (Research England, the organisation which organises the REF, destroyed all unit-level scores after the completion of the exercise, Brisson, 2024). In other words, we used ecological correlations: the correlations between two group means (unit-level output GPAs and topic weightings of units' composite mean papers). Usually, ecological correlations are stronger than the equivalent correlations calculated using individual-level data (e.g. Robinson, 2009), and assuming that individual- and group-level correlations are equivalent is sometimes referred to as the ecological fallacy. The fallacy can be demonstrated by comparing the group-level correlation between citation counts and REF2014 quality judgements reported by Pride and Knoth (2018) and the output-level correlation between the same variables reported by Wilsdon et al. (2015) in their REF-commissioned study of whether metrics could replace expert peer review in the REF. For the psychology, psychiatry and neuroscience subpanel, Pride and Knoth found a correlation of .651 between units' mean citation counts and their output GPAs, whereas Wilsdon et al. (who had access to the output-level scores) found an individual-level correlation of .407. An analogous reduction of the .794 correlation we found between predicted REF2014 output GPAs and actual REF2014 output GPAs might be expected if we were able to conduct our analysis at the output level rather than the unit level, although estimating the size of any such reduction with accuracy is impossible. This suggests that drawing strong conclusions about the quality of any individual output based on the kind of model we have offered here would be unwise.

Evidence of bias

Was there 'a problematic bias in favour of biological and cognitive psychology' within the REF2021 psychology, psychiatry and neuroscience panel, as suggested by Langdridge (2020)? Possibly. As shown in Figure 4, our data robustly demonstrate that submissions which included a great deal of qualitative research received lower scores than those which returned little or no qualitative research. This was true in both REF2021 and REF2014. The zero-order correlation between the units' proportions of words from the Qualitative Research topic and their output GPAs in REF2021 was $r = -0.796$. This correlation remained strong after controlling for the full-time equivalent number of staff (FTE) returned by each unit, and the amount of grant income from UK Research and Innovation, the British Academy and the Royal Society spent per FTE by each unit, $pr = -0.763$. These correlations are not driven by the small number of outliers

with extremely high proportions of words from the qualitative research topic. Restricting the REF2021 analysis to those 80 units in which qualitative research made up less than 20% of the composite mean paper yielded a strong correlation of $r = -0.639$.

Before considering possible accounts for this finding, we note that our data are not consistent with a general bias in favour of biological and cognitive psychology, as Langdridge also suggested. Some cognitive topics, notably face recognition, were also associated with lower scores, as were some more biologically based topics, such as psychopharmacology.

What might lie behind the strong relationship we observed between units returning more qualitative research and receiving lower output quality assessments? There are at least four possibilities, not mutually exclusive, which might account for these data.

One account is that the subpanel found it hard to fairly and accurately assess the quality of qualitative research, which made work of this sort less likely to receive the highest scores. Considerations regarding the structure of academic disciplines from the philosophy of science literature suggest that this is plausible. Consider, for instance, Lakatos (1978) methodology of scientific research programmes. Lakatos argued that a scientific discipline, such as psychology, is made up of rival research programmes that adopt different 'hard cores': assumptions and beliefs that are accepted by all those who work within the programme. Because of differences in these hard cores, research programmes also typically adopt different 'heuristics', the collection of methods and problem-solving techniques used to make progress within the programme. For instance, measuring response times is an important part of the heuristic for the cognitive psychology programme, as a core assumption of the programme's hard core is that 'thinking' can be conceptualised as the processing of information, and because more complex processing is assumed to take longer than less complex processing. However, a researcher working within a research programme where 'thinking' is conceptualised not as information processing, but rather as participation within a social or cultural practice, is very unlikely to be interested in measuring response times.

What happens when a researcher who works within one research programme is asked to evaluate the quality of a piece of research from another programme, particularly a piece from a less prominent programme? As Gillies (2008) has argued, researchers inevitably tend to be more sympathetic to approaches from research programmes that are like the programme that they themselves favour. This, coupled with a likely lack of familiarity with the assumptions embedded in the hard cores of minority research programmes is likely to create a conservative bias in the peer review process in favour of majority research programmes. This is particularly likely to be the case when minority programmes are radically different from dominant programmes, for instance in the case of Kuhnian paradigm shifts (Gillies, 2008).

Does this apply to the case of qualitative research in the psychology, psychiatry and neuroscience subpanel? We believe it is likely. 'Qualitative research' includes a wide range of research programmes,

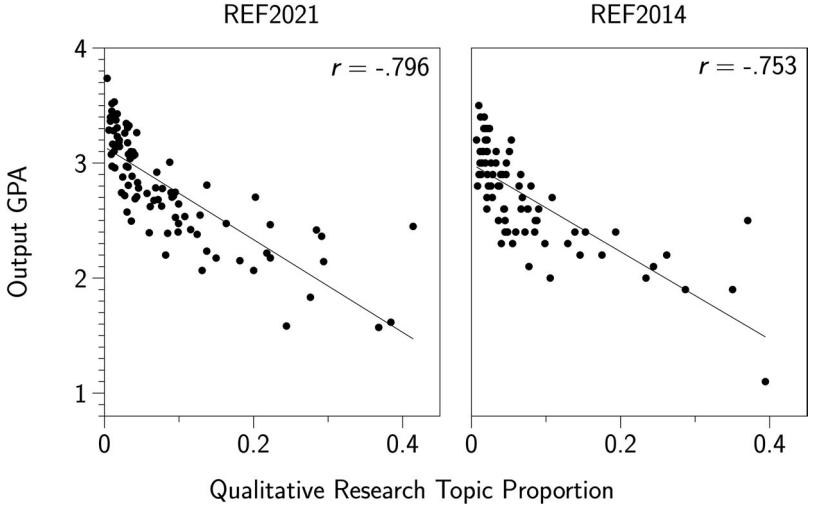

Figure 4. Plots showing the relationship between the proportion of units' composite mean papers that came from the Qualitative Research topic and their output GPAs from both REF2014 and REF2021.

some of which are antithetical to one another. But many (although not all) of these programmes adopt quite different epistemological assumptions to those where quantitative methods are more common. Moreover, this may not be well understood by researchers inexperienced in these programmes. Some empirical evidence supports this suggestion: Clarke et al. (2025) analysed the peer review experiences of 163 qualitative researchers, finding that many reported experiencing 'methodologically incongruent reviewing', usually in the form of reviewers who assumed that the perspectives and standards associated with typical quantitative research programmes are universally applicable. Most subpanel members in 2021 were quantitative experts, to the extent that in response to concerns from the research community a researcher with expertise in qualitative research (Brendan Gough) was added after the first round of panel membership announcements in 2018 (BPS, 2019). These factors suggest that the mechanism discussed by Gillies (2008) may well have been present, leading to lower scores, on average, for the qualitative research returned to the subpanel.

A second possibility is that the subpanel was perceived as being likely to assess qualitative research negatively by those colleagues responsible for selecting outputs for submissions, particularly in units where there was a strong pool of outputs from which to select. If the panel was perceived as being less likely to award high scores to qualitative research (regardless of whether this perception was accurate), then selecting qualitative papers might be considered a risk. In departments where this risk could be mitigated by selecting other work, it presumably would have been. If those departments where there was a surplus of high-quality non-qualitative research were also those where higher quality research is done in general, then this mechanism would create an artefactual relationship between the amount of qualitative work returned and output quality. Nevertheless, recall that the correlation between the amount of qualitative work a unit returned and their output GPA remained strong even after controlling for the number of academics and the amount of grant income per academic spent by the unit. If these two variables track the number of high-quality outputs a unit had to choose from, this account probably cannot solely be responsible for the associations we observed.

A third possibility is that the panel was perceived as being likely to assess qualitative research negatively by those colleagues responsible for deciding which units a university should return to. If those departments where high-quality qualitative psychology is conducted were more likely to return their psychology researchers to subpanels other than psychology, psychiatry and neuroscience, then again this could account for our observed relationship. There is some independent evidence consistent with such a possibility. As noted above, only 78.5% of academic psychologists surveyed by the BPS (2014) reported that they were returned to the psychology, psychiatry and neuroscience subpanel in REF2014; and, at the time of writing, there were 33 universities who delivered psychology programmes accredited by the BPS who did not make returns to the subpanel (compared to the 89 accredited universities who did, and four who made a return but who did not have accredited courses). At least some of these universities enjoy a reputation for high-quality qualitative psychology research (e.g. Loughborough, as described by Stokoe, Hepburn & Antaki, 2012).

A final possibility is that qualitative psychology research is in fact, on average, of lower quality than quantitative psychology research, in terms of originality, significance and rigour. Yet there is ample independent evidence to support Gathercole's assertion that many UK researchers excel in qualitative psychology (Brooks et al., 2018, p. 5). For instance, between 2013 and 2021, sixteen psychologists were awarded Honorary Fellowships of the BPS, with at least five of these colleagues being primarily known for their qualitative work. Similarly, several qualitative-focused psychology journals have very high source-normalised impact factors (SNIP; Moed, 2010), demonstrating that some qualitative research does attract considerable attention (e.g. *Qualitative Research in Psychology* has, at the time of writing, the second highest SNIP of all psychology journals, and *Qualitative Psychology* has the eighth highest).

However, the existence of high-quality (or at least high-visibility) qualitative psychology research does not rule out the possibility that qualitative research is, on average, 'worse' than quantitative research. Would such a claim be meaningful? And if it were meaningful, would it be plausible? To answer this question, we can again appeal to Lakatos's (1978) methodology of scientific research programmes. He noted that research programmes can be categorised as either 'progressing' or 'degenerating'. Progressing

programmes are those which use their heuristics to regularly find surprising and important new results. Degenerating programmes are those which rarely produce novel insights and instead tend to focus on ad hoc accommodations to deal with anomalous observations. Eventually, Lakatos suggested, researchers working within a degenerating programme will notice this, and abandon it. Might qualitative psychology be conceptualised as consisting of degenerate research programmes in this sense? If so, this might provide some justification for the relationships we observed. However, given Gathercole's observation, given the number of qualitative psychologists who have been honoured by the BPS for the quality of their work and given the existence of thriving qualitatively focused psychology journals, the suggestion that qualitative psychology is degenerate in Lakatos's sense seems an implausible account for our findings.

Although we cannot conclusively distinguish between these four accounts of the relationship between submitting more qualitative research and receiving lower quality profiles, encouraging a wider discussion of these possibilities is important. The first three accounts, if true, would clearly indicate a problem either with the processes used in research assessment in our discipline, or with how those processes are perceived. If one or more of these accounts is correct, then taking mitigating action to ensure that the REF welcomes the full range of psychology research seems necessary.

Notes

1. Although how to characterise academic disciplines and their boundaries is not straightforward (e.g. Becher & Trowler, 2001; Bridges, 2006).
2. Although there are also political reasons to suggest that the discipline might benefit from this organisation. The mechanism by which REF results are translated into funding depends, in part, on the cost bands assigned to each subject. Since 2019/20 psychology, psychiatry and neuroscience has been deemed a band A subject (the highest, characterised as 'high-cost laboratory and clinical subjects'). Some colleagues believe that psychology being included in a single subpanel alongside psychiatry and neuroscience makes it easier to justify this banding.
3. Where a particular paper was returned by one or more units (perhaps it had coauthors from several institutions), these instances were treated independently. Because MALLET uses Gibbs sampling, a stochastic process, these duplicates should be expected to have very slightly different topic proportions.
4. According to the Stern (2016) review, the REF has multiple purposes. These include both allocating funding and providing reputational benchmarking information regarding research quality. Because the funding associated with the REF is disproportionately affected by 4* scores, arguably using GPA as a dependent variable is less useful if funding is considered to be the REF's primary purpose. Here we were more concerned with assessing research quality, which is better indexed by GPA, as this is the typical measure used in newspaper league tables (such as those published by the Times Higher Education). Relatedly, some universities favour reporting 'research power' rather than GPA, which is calculated by multiplying a unit's GPA by the FTE number of staff they returned. 'Research power' correlates extremely strongly with FTE ($r > 0.99$), and we do not consider it to be a useful index of research quality.

Ethical approval

There are no ethical issues raised.

Author contributions

Conceptualisation: Matthew Inglis, Elizabeth Stokoe. Formal analysis: Matthew Inglis, Colin Foster, Hugo Lortie-Forgues, Victoria Simms, Elizabeth Stokoe. Writing – original draft: Matthew Inglis. Writing – review & editing: Matthew Inglis, Colin Foster, Hugo Lortie-Forgues, Victoria Simms, Elizabeth Stokoe.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by Research England, via an Expanding Excellence in England grant to the Centre for Mathematical Cognition, and the Economic and Social Research Council [grant number ES/W002914/1].

ORCID

Matthew Inglis http://orcid.org/0000-0001-7617-4689
Colin Foster http://orcid.org/0000-0003-1648-7485
Hugues Lortie-Forgues http://orcid.org/0000-0002-4060-8980
Victoria Simms http://orcid.org/0000-0001-5664-6810
Elizabeth Stokoe http://orcid.org/0000-0002-7353-4121

Data availability statement

Online materials associated with this manuscript are available at https://doi.org/10.17028/rd.lboro.28881821. Data and analysis scripts associated with this manuscript are available at http://doi.org/10.17028/rd.lboro.28881821

References

Becher, T., & Trowler, P. (2001). *Academic tribes and territories*. McGraw-Hill Education.
Bence, V., & Oppenheim, C. (2005). The evolution of the UK's research assessment exercise: Publications, performance and perceptions. *Journal of Educational Administration and History*, *37*(2), 137–155. https://doi.org/10.1080/00220620500211189
Berk, R. A. (2004). *Regression analysis: A constructive critique*. Sage.
Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *82*, 48–62. https://doi.org/10.1016/j.cortex.2016.05.003
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.
Bridges, D. (2006). The disciplines and discipline of educational research. *Journal of Philosophy of Education*, *40*(2), 259–272. https://doi.org/10.1111/j.1467-9752.2006.00503.x
Brisson, R. (2024). REF 2029 data 'could be made more available for research'. Research Professional News. https://www.researchprofessionalnews.com/rr-news-uk-research-councils-2024-3-ref-2029-data-could-be-made-more-available-for-research/.
British Psychological Society. (2019). Research *Excellence Framework* 2021. https://www.bps.org.uk/blog/research-excellence-framework-2021
Brooks, J., Goodman, S., Locke, A., Reavey, P., Riley, S., & Seymour-Smith, S. (2018). *Writing for the Research Excellence Framework 2021: Guidance for qualitative psychologists*. British Psychological Society.
Clements, C., Hawton, K., Geulayov, G., Waters, K., Ness, J., Rehman, M., Townsend, E., Appleby, L., & Kapur, N. (2019). Self-harm in midlife: Analysis using data from the multicentre study of self-harm in England. *The British Journal of Psychiatry*, *215*(4), 1–8. https://doi.org/10.1192/bjp.2019.90
Coenders, G., & Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, *44*(1), 201–220.
Collins, A., & Bunn, G. (2016). The shackles of practice: History of psychology, research assessment, and the curriculum. In S. H. Klempe & R. Smith (Eds.) *Centrality of history for theory construction in psychology* (pp. 91–109). Springer.
Clarke, V., Braun, V., Adams, J., Callaghan, J. E., LaMarre, A., & Semlyen, J. (2025). "Being really confidently wrong": Qualitative researchers' experiences of methodologically incongruent peer review feedback. *Qualitative Psychology*, *12*(1), 7–24. https://doi.org/10.1037/qup0000322
Gillies, D. (2008). *How should research be organised?* College Publications.
Inglis, M., Foster, C., Lortie-Forgues, H., & Stokoe, E. (2024). British education research and its quality: An analysis of Research Excellence Framework submissions. *British Educational Research Journal*, *50*(5), 2495–2518. https://doi.org/10.1002/berj.4040
Jacobi, C., Van Atteveldt, W., & Welbers, K. (2018). Quantitative analysis of large amounts of journalistic texts using topic modelling. In M. Karlsson & H. Sjøvaag (Eds.) *Rethinking research methods in an age of digital journalism* (pp. 89–106). Routledge.
Jones, P., & Sizer, J. (1990). The universities funding council's 1989 research selectivity exercise. *Beiträge Zur Hochschulforschung*, *4*, 309–348.
Koya, K., & Chowdhury, G. (2017). Metric-based vs peer-reviewed evaluation of a research output: Lesson learnt from UK's national research assessment exercise. *PLOS One*, *12*(7), e0179722. https://doi.org/10.1371/journal.pone.0179722
Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers* (Vol. 1). Cambridge University Press.
Langdridge, D. (2020). *REF 2021 – What it is and why it matters* [Blogpost]. https://oupsychology.wordpress.com/2020/09/07/ref-2021-what-it-is-and-why-it-matters/ Retrieved on 12/2/25.

Lewis, S., Anderson, D., Lyonette, C., Payne, N., & Wood, S. (2017). Public sector austerity cuts in Britain and the changing discourse of work–life balance. *Work, Employment and Society*, *31*(4), 586–604. https://doi.org/10.1177/0950017016638994

Marques, M., Powell, J. J., Zapp, M., & Biesta, G. (2017). How does research evaluation impact educational research? Exploring intended and unintended consequences of research assessment in the United Kingdom, 1986–2014. *European Educational Research Journal*, *16*(6), 820–842. https://doi.org/10.1177/1474904117730159

McCallum, A. K. (2002). MALLET: MAchine Learning for LanguagE Toolkit. Retrieved from http://mallet.cs.umass.edu

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Munoz-Chereau, B., & Wyse, D. (2023). *Education: The state of the discipline. The progress of education an analysis of data from the Research Excellence Framework*. British Educational Research Association. https://www.bera.ac.uk/publication/education-the-state-of-the- discipline-progress-of-education

Pride, D., & Knoth, P. (2018). Peer review and citation data in predicting university rankings, a large-scale analysis. In E. Méndez, F. Crestani, C. Ribeiro, G., & David, J. Lopes (Eds.) *Digital libraries for open knowledge*. TPDL 2018. *Lecture notes in computer science*, vol 11057. Springer. https://doi.org/10.1007/978-3-030-00066-0_17

Poppler [Computer Software] (2022). Retrieved from https://poppler.freedesktop.org.

Rocchi, F., Ledgeway, T., & Webb, B. S. (2018). Criterion-free measurement of motion transparency perception at different speeds. *Journal of Vision*, *18*(4), 5. https://doi.org/10.1167/18.4.5

REF. (2022). Overview *report by main panel* A and *sub*-panels 1 to 6. https://2021.ref.ac.uk/media/1910/mp-a-overview-report-final-updated-september-2022.pdf.

Research Board of the British Psychological Society. (2014). *Report on the outcomes of the REF experiences survey*. Unpublished Internal Report, British Psychological Society.

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, *38*(2), 337–341. https://doi.org/10.1093/ije/dyn357

Royal Geographical Society. (2017). Consultation *on the second Research Excellence Framework*. https://www.rgs.org/about-us/what-is-geography/consultations/second-research-excellence-framework-ref2021

Simms, V., Gilmore, C., Cragg, L., Clayton, S., Marlow, N., & Johnson, S. (2015). Nature and origins of mathematics difficulties in very preterm children: A different etiology than developmental dyscalculia. *Pediatric Research*, *77*(2), 389–395. https://doi.org/10.1038/pr.2014.184

Stern, N. (2016). Research Excellence Framework review. UK Government. https://www.gov.uk/government/publications/research-excellence-framework-review [accessed 10 September 2025].

Stokoe, E., Hepburn, A., & Antaki, C. (2012). Beware the 'Loughborough School' of social psychology? Interaction and the politics of intervention. *The British Journal of Social Psychology*, *51*(3), 486–496. https://doi.org/10.1111/j.2044-8309.2011.02088.x

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Vine, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. HEFCE.

Wetherell, M. (2011). The winds of change: Some challenges in reconfiguring social psychology for the future. *The British Journal of Social Psychology*, *50*(3), 399–404. https://doi.org/10.1111/j.2044-8309.2011.02038.x