Received: 13 February 2024 | Accepted: 20 May 2024

DOI: 10.1002/berj.4040

BERJ BERA

ORIGINAL ARTICLE

British education research and its quality: An analysis of Research Excellence Framework submissions

Matthew Inglis[1] | Colin Foster[1] | Hugues Lortie-Forgues[1] | Elizabeth Stokoe[2]

[1]Centre for Mathematical Cognition, Loughborough University, Loughborough, UK

[2]Department of Psychological and Behavioural Science, London School of Economics and Political Science, London, UK

Correspondence
Matthew Inglis, Centre for Mathematical Cognition, Loughborough University, Loughborough LE11 3TU, UK.
Email: m.j.inglis@lboro.ac.uk

Funding information
Economic and Social Research Council; Research England

Abstract

We analysed the full text of all journal articles returned to the education subpanel of the 2021 Research Excellence Framework (REF2021). Using a latent Dirichlet allocation topic model, we identified 35 topics that collectively summarise the journal articles that research units, typically schools of education, selected for submission. We found that the topics which units wrote about in their submitted articles collectively explained a large proportion (84.1%) of the variance in the quality assessments they received from the REF's expert peer review process. Further, with the important caveat that we cannot attribute causality, we found that there were strong associations between what the subpanel perceived to be excellent research and the adoption of particular methods or approaches. Most notably, units that returned more interview-based work typically received lower scores, and those which returned more analyses of large-scale data and meta-analyses typically received higher scores. Finally, we applied our 2021 model to articles submitted to the previous exercise, REF2014. We found that education research seems to have become less qualitative and more quantitative over time, and that our 2021 model could successfully predict the scores assigned by the REF2014 subpanel, suggesting a reasonable degree of between-exercise consistency.

KEYWORDS

assessment, excellence, quality, research, research approaches

Key insights

What is the main issue that the paper addresses?

We analysed all the papers submitted to the education subpanel of REF2021 with the aim of understanding (i) the topics that British education research focuses upon and (ii) whether the topics that units write about predicts the quality assessments they receive from the REF's expert peer review process.

What are the main insights that the paper provides?

The topics which units wrote about in their submitted articles collectively explained a large proportion (84.1%) of the variance in the quality assessments they received from the REF's expert peer review process. Moreover, our model from REF2021 was also successful at predicting quality judgements made in REF2014.

THE RESEARCH EXCELLENCE FRAMEWORK

The UK government's research funding bodies periodically conduct an assessment of the research quality of each publicly funded university, with the aim of informing how to allocate ‘quality-related’ research funding. These assessment exercises are extremely important for academics and universities in the United Kingdom, because they influence both funding and reputation. As a result, the process by which research quality judgements are made during such exercises is of considerable interest to UK-based academics. However, understanding this process is not merely of parochial British interest. Many other countries have adopted research evaluation systems that share several characteristics of the UK system (e.g., Geuna & Martin, 2003; Pinar & Horne, 2022), and even in contexts where this is not the case, the assessment of research quality is central to the process of appointing and promoting academics. Our goal is to systematically interrogate how judgements of research quality are made.

The most recent of the United Kingdom's periodic research assessments, known as the Research Excellence Framework (REF), took place in 2021. The exercise involves groups of academics being submitted as ‘units’ to assessment panels (or, in REF terminology, subpanels) defined by discipline, such as education. So, in a typical UK university, academics working in the School of Education were collectively submitted as a unit to the education subpanel. The rules on who can/must be included in the assessment have varied over time, as have the rules on the number of research outputs (the generic term for journal articles, books, chapters, conference proceedings, etc.) that can be submitted. In REF2021, each researcher was required to submit between one and five outputs, with each unit as a whole submitting an average of 2.5 outputs per researcher.

Once submitted, these research outputs were assessed for their quality by a panel of senior academics and other experts. In the education case, the REF2021 subpanel consisted of 22 full members, all senior academics from the field, whose remit was to develop the assessment criteria and conduct assessments, together with a further 14 assessors who only participated in the assessment phase.

Each output submitted to the education subpanel was given a score for its quality in terms of ‘originality, significance and rigour’ on a five-point scale: from the highest 4* rating to unclassified, as shown in Table 1. These scores were combined to produce an output quality profile for each unit, which contributed to the overall quality profile, alongside analogous profiles for the reach and significance of the unit's impact (assessed via case studies; REF, 2019) and the extent to which the unit's environment is conducive to producing

TABLE 1 The five possible REF quality ratings.

Rating	Description
4*	Quality that is world-leading in terms of originality, significance and rigour
3*	Quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence
2*	Quality that is recognised internationally in terms of originality, significance and rigour
1*	Quality that is recognised nationally in terms of originality, significance and rigour
u/c	Quality that falls below the standard of nationally recognised work, or work which does not meet the published definition of research for the purposes of this assessment

high-quality research (assessed via narrative statements and various metrics; REF, 2019). One convenient way of expressing REF quality profiles is to calculate a grade point average (GPA). For instance, in REF2021 the outputs in the University of Aberdeen's education submission received a quality profile of 22.5% 4*, 30.0% 3*, 40.0% 2*, 5.0% 1* and 2.5% unclassified. So, their output GPA was 2.35 ($0.225 \times 4 + 0.3 \times 3 + 0.4 \times 2 + 0.05 \times 1 + 0.025 \times 0$).

The outcomes of REF exercises determine the amount of government research funding each institution receives and influence their reputations (e.g., REF scores contribute to some domestic newspaper league tables). As a result, REF exercises are taken remarkably seriously: much effort is devoted to selecting the outputs that are deemed most likely to receive high grades. Partly as a result of this high-stakes nature, the exercise has received a great deal of academic attention. It has been criticised in terms of its underpinning political stance (e.g., Brown & Carasso, 2013; Fairclough, 1995), of the impact that it has on interdisciplinary research (e.g., Pardo-Guerra, 2022) and of its unintended consequences (e.g., Brassington, 2022; Gillies, 2008; Marques et al., 2017; Pinar & Horne, 2022; Watermeyer & Derrick, 2022).

There are at least three reasons why education researchers should be interested in understanding both what was submitted to the education panel in REF2021 and how these submissions were assessed. First, submissions to the REF provide a snapshot of the state of each UK university's most highly regarded education research in the period 2014–2021 (at least the most highly regarded, subject to the rules of the exercise). By analysing the papers that were submitted, we can begin to understand the topics UK-based researchers focused on, the theoretical perspectives and methodological approaches most commonly used, and where there are gaps. Second, the REF panel was made up of highly respected academics who were tasked with peer reviewing a large number of education research outputs with the goal of producing a careful estimate of their quality. Understanding the factors that influenced their judgements gives us insights into the strengths and weaknesses of contemporary British education research, or at least the factors that influence expert peer review. Third, while the rules of future REF exercises will evolve, the centrality of an expert peer review assessment of output quality within disciplines is likely to remain. Given this, it is important to explore between-panel consistency of judgement. Does the type of research that is assessed as being high quality remain stable across years, or does it vary from exercise to exercise?

The purpose of the current paper is to report a study that analysed the full text of all journal articles submitted to the REF2021 education subpanel. Our specific goals were to identify the makeup of these papers in terms of their substantive content, their methodological focus and their theoretical orientation, and then determine whether these factors were associated with the output quality assessments made by the panel. In the last stage, we applied our model to the full text of journal articles submitted to the previous REF (REF2014) and assessed (i) changes in substantive content of REF papers over time and (ii) between-panel stability of quality judgements. Before reporting the steps we used in our analysis, we first introduce the method we used: latent Dirichlet allocation topic modelling.

TOPIC MODELLING THE RESEARCH EXCELLENCE FRAMEWORK

Topic modelling is a method that seeks to understand the content of a large number of texts by identifying the themes, or topics, that they focus on (Blei et al., 2003). The method takes a large collection of unstructured texts and studies the words they contain. For instance, if a document contains many instances of the words 'sofa', 'table' and 'armchair', we might infer that the document is, to some extent at least, about furniture. Formally, a topic is defined to be a probability distribution over words. So, a furniture topic would associate high probabilities to words related to furniture ('sofa', 'table', 'armchair') and low probabilities to words unrelated to furniture ('biscuit', 'fishing', 'stockbroker').

To understand topic modelling it is helpful to consider the process in reverse, where we imagine that we have a collection of topics and want to create some documents. Imagine that we want to write a document that is 50% about furniture, 30% about silent films and 20% about Nottingham. Every time we want to add a word to our document, we select it from the furniture topic with probability 0.5, from the silent films topic with probability 0.3 and from the Nottingham topic with probability 0.2. Each topic is itself a set of probabilities over words; for instance, perhaps the Nottingham topic assigns a probability of 0.01 to 'Trent'. So, we know that every time we want to add a word to our document, the probability of it being 'Trent' (from the Nottingham topic) is 0.2×0.01. This method of creating documents involves two considerable simplifications. First, the so-called 'bag of words' model of text is adopted by ignoring word order. Second, '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 what the most likely topics were. After identifying these topics, the composition of each document can 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).

Topic modelling has previously been used to shed light on a variety of issues in educational research. Inglis and Foster (2018) topic modelled the full texts of all articles published in the two leading mathematics education journals, the *Journal for Research in Mathematics Education* and *Educational Studies in Mathematics*, since their foundation. They found topics associated with the specific content of research (e.g., spatial reasoning, algebra, etc.) but also with theoretical perspectives (e.g., constructivism, sociocultural theories, etc.). Using the compositions of each published paper, Inglis and Foster were able to quantitatively track the changing theoretical orientations adopted by mathematics education researchers since the 1970s. Using a similar method, Galvez et al. (2019) analysed the abstracts of 137,000 dissertations completed between 1980 and 2010 in the United States. They particularly focused on the relative prominence of what they referred to as the interpretative and causal research paradigms, finding that the prominence of the interpretative paradigm had increased during this period, whereas the prominence of the causal paradigm had declined.

Both Inglis and Foster (2018) and Galvez et al. (2019) highlighted that the (quantitative) topic modelling method and qualitative methods like grounded theory share a focus on bottom-up inductive coding of the data. The topic modeller has no preconceived expectation about what topics will emerge from their unstructured collection of documents, and they must interpret the meaning of those topics that do emerge. These interpretations must, as with a grounded theory analysis, be carefully justified. Later in the paper we explain the methods we adopted in the current project, and justify our interpretations of the topics that emerged. However, we first review what we know about the REF2021 education subpanel's judgements of output quality.

REF2021 EDUCATION OUTPUT QUALITY

The quality assessments given to each individual output by REF panels are confidential, but the aggregated quality profiles for each submitted unit are published online at www.ref.ac.uk. This, along with the report written by the education subpanel after the results were published, provides us with the best source of evidence about how the panel went about judging papers. Comparing the quality profiles awarded by the education subpanel to other disciplinary subpanels reveals that the education panel generally judged the quality of the outputs it assessed to be rather weak. Figure 1 shows the distribution of profiles awarded by each subpanel, and indicates that the education subpanel awarded high numbers of 1* and 2* ratings and low numbers of 3* and 4* ratings, at least in comparison to the other disciplinary subpanels.

How did the subpanel reach its judgements? The post-assessment subpanel report indicated that ‘detailed calibration exercises were conducted’ and that ‘processes for moderation were used throughout’, which ‘included paired assessment, monitoring of scoring patterns from the subpanel (individually and collectively) and from the main panel’ (REF, 2022, p. 158). In its summary of the output assessment process, the subpanel noted that outputs could achieve the highest grades ‘in diverse ways’ and that there was ‘no strong association between research excellence and particular methods or approaches’ (p. 159). Furthermore, the report noted that outputs awarded low grades typically exaggerated their contributions to knowledge, were poorly situated within a field, offered insufficient justification of their sampling strategy, or had underdeveloped ‘criticality and analytical purchase’ (p. 159).

The remainder of the subpanel's report highlighted particular areas of strength of the submitted research. For instance, there was ‘especially strong work on the identities of children and young people, focused on gender, sexuality, race, ethnicity and socioeconomic background’ (p. 162) and ‘educational research drawing on philosophy and history was mainly of very high quality’ (p. 164). In terms of methodological focus, the subpanel remarked that designs focused on the ‘analysis of qualitative data remain common in education and much of this work continued to be of the highest quality’ (p. 165) and that there were ‘strong examples of the use of longitudinal data to track long-term outcomes in education’ (p. 165). Because the methodological strengths noted were specific examples rather than general trends (e.g., the subpanel stated that there were ‘strong examples’ of longitudinal studies,

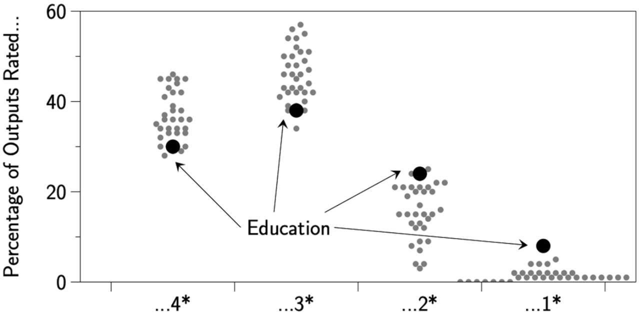

FIGURE 1 The percentage of outputs rated 4*, 3*, 2* and 1* for each disciplinary subpanel (where each dot represents a subpanel), with the education subpanel's figures highlighted.

rather than making the stronger claim that longitudinal studies tended to be, on average, high quality), the existence of these strengths did not contradict the most important assertion in the subpanel report, that there was 'no strong association' between quality judgements and research methods or approaches (p. 159).

METHOD

In total, 5272 outputs were submitted to the education subpanel of REF2021. Of these, 4295 (81.5%) were identified as journal articles by the submitting units. We obtained pdf copies of 4290 of these articles (the remaining five were identified by the submitting units as being written in a language other than English, and so would not have been analysable with an English-language topic model). We converted these 4290 pdfs to plain text using the UNIX pdftotext command (Poppler, 2022) and then used MALLET (version 2.0.8RC2), a UNIX topic modelling tool (McCallum, 2002), to calculate possible models, ignoring the stop words on MALLET's default list.

To evaluate the optimal number of topics for our main model, we adopted the perplexity approach (Blei et al., 2003; Jacobi et al., 2018). We split the corpus into a training corpus (80%) and a testing corpus (20%), fitted topic models with 10 topics, 20 topics, … 100 topics to the training corpus and then calculated the perplexity of each using the testing corpus. Perplexity is an estimate of model fit, with a lower value indicating a better fit. Perplexity values for each model are shown in Figure 2. Clearly, choosing a model with more topics will lead to a lower perplexity figure, and Jacobi et al. suggested choosing the number of topics based on where the relationship between perplexity and topic numbers 'levels off' (much like a scree plot in an exploratory factor analysis). Based on the piecewise linear regression shown in Figure 1, we opted to fit a model with 35 topics for our main analysis.

This allowed us to calculate the composition of each of the 4290 English-language journal articles returned to the education subpanel. For example, consider Bennett et al.'s (2019) article entitled 'The cost of multiple representations: Learning number symbols with abstract and concrete representations', published in the *Journal of Educational Psychology*. The article explored whether children's learning of number symbol meanings varied depending on the type of representation they were introduced with (abstract or concrete). Our 35-topic model identified that 68.2% of the article's words came from Topic 5 and 8.6% from Topic 7 (note that 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 Developmental Psychology and Mathematics, respectively, which seems to capture the content of Bennett et al.'s article well.

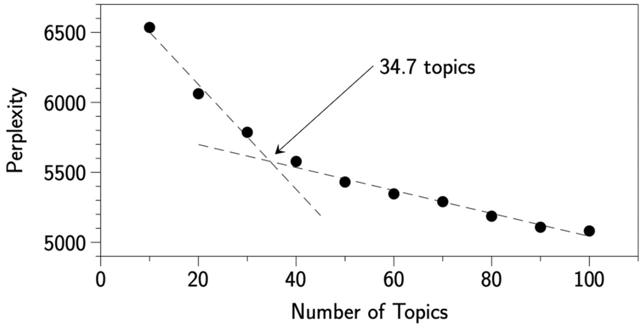

FIGURE 2 Perplexities associated with models with 10, 20, 30, …, 100 topics. The dotted lines show a one-break piecewise linear regression line of best fit.

RESULTS

The defining words associated with the 35 topics identified in our model are shown in Table 2, together with the article that had the highest proportion of words from the topic, the name we gave to describe the topic and the mean proportion of words from each topic (averaged over all papers). These names were assigned based on the defining words and, where that was not sufficient, a careful reading of papers with the highest proportions of words from the topic and papers with the lowest proportions of words from the topic. In most cases it was straightforward to assign names to topics. For example, the topic characterised by words such as 'gender', 'girls', 'boys', 'male', 'female', 'men' and 'identity' was clearly about gender (readers can assess the adequacy of our names by consulting Table 2, and the online data associated with this paper, available at https://doi.org/10.17028/rd.lboro.25201139.v1). There were three exceptions—Topics 3, 11 and 20—where a careful reading of outputs was required.

Topic 3 was characterised by words such as 'research', 'assessment', 'data', 'analysis', 'evidence', 'review', 'knowledge' and 'approach'. The papers with the highest proportions of words from the topic were all focused on developing and/or evaluating methodological approaches. For instance, Hitchcock and Onwuegbuzie's (2020) article 'Developing mixed methods crossover analysis approaches' was made up of 68.1% of words from the topic, and Nelson's (2017) discussion of criteria for saturation in qualitative research contained 65.2% of words from the topic. The topic seemed to be focused on the discussion of methods per se, including their innovation, systematic application and rigorous use. The topic did not focus on a particular type of method: both quantitative and qualitative approaches were discussed by papers that had high proportions of words from the topic (e.g., McGrane et al., 2018; Sechelski & Onwuegbuzie, 2019). We named the topic Methodological Depth.

Topic 11 was characterised by words such as 'school', 'schools', 'education', 'del', 'und', 'Italian', 'della', 'der', 'les' and 'educacion'. Upon reading papers with particularly high proportions of words from this topic, we concluded that the topic captured outputs that included passages of text not written in English. For instance, perhaps the output had two abstracts—one written in English, one in a non-English language (e.g., Myhill & Jones, 2015). We named this topic Non-English Components.

Topic 20 was characterised by words such as 'first', 'significant', 'findings', 'specific', 'influence', 'field', 'effects', 'find', 'differences', 'reflect' and 'significantly'. Unlike most other topics, there were few articles that had particularly high proportions of words from the topic. The largest was Baird et al.'s (2017) article 'Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems', which had 44% of its words from Topic 20. This article was notable for the emphasis made on asserting the originality and significance of the reported research. For instance, Baird et al. noted that their 'study is the first to show instability across monitoring systems' (p. 11), that it is 'the first study to show this [result] as a general effect, rather than for a particular team, and [that] it is the first to use multilevel modelling to do so' (p. 11). Furthermore, they argued that their results are 'important findings, as face-to-face training and Supervisor-based monitoring systems are still the norm in many examination settings for practical reasons' (p. 12). In short, the paper attempted to make a particularly strong case for the originality and wider significance of its findings. No other paper had nearly as high a proportion of words from Topic 20 (the next highest was 23%), but all the papers with proportions over 20% also discussed the significance of their findings (e.g., they developed a wider theoretical framework or discussed the implications of their results for practice at length). For example, Gibbs and Elliott's (2015) study of how teachers interpret terms such as 'dyslexia' described how their findings 'provide a potential challenge to the value, meaning and impact of certain labels that may be used as "short-hand" descriptors for the difficulties that some children experience' (p. 335). In contrast, when we studied

TABLE 2 The 35 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 number	Topic name	Characteristic words	REF2021 mean %	REF2014 mean %	Paper with highest proportion of words from the topic
1	Communication and Interaction	Writing interaction talk text language analysis discourse texts classroom communication dialogue understanding dialogic students social written meaning writers interactions literacy	2.32	2.72	Ingram, J., & Elliott, V. (2014). Turn taking and ‘wait time’ in classroom interactions. *Journal of Pragmatics*, *62*, 1–12
2	Philosophy of Education	Education human knowledge world philosophy educational moral theory university press work critical society view sense life thinking form idea political	5.75	6.94	Morgan, M. (2016). Hannah Arendt and the ‘freedom’ to think. *Journal of Educational Administration and History*, *48*(2), 173–182
3	Methodological Depth	Research assessment data studies analysis evidence methods review knowledge process approach practice researchers quality evaluation development impact e.g. framework study	7.99	8.70	Hitchcock, J. H., & Onwuegbuzie, A. J. (2020). Developing mixed methods crossover analysis approaches. *Journal of Mixed Methods Research*, *14*(1), 63–83
4	Citizenship and Culture	Political citizenship values Ireland rights northern contact British community European intercultural identity people prevent Muslim national migration social civic attitudes	1.40	1.14	Hodgkin, K., Bethell, S., Bryant, A. S., Edwards, L. C., & Cooper, S. M. (2020). Mentoring PE student teachers in Wales: Lessons from a systematic review of the literature. *Wales Journal of Education*, *22*(2), 26–51
5	Developmental Psychology	Children number task development children' age child psychology trials model participants object numbers developmental cognitive study effect e.g. tasks cognition	1.45	1.30	Sella, F., & Lucangeli, D. (2020). The knowledge of the preceding number reveals a mature understanding of the number sequence. *Cognition*, *194*, 104104
6	Health and Medicine	Health medical care study clinical patient patients doctors training research journals staff med nursing professional reproduction work healthcare medicine NIHR	1.51	1.22	Duley, L., Dorling, J., Ayers, S., Oliver, S., Yoxall, C. W., Weeks, A., et al. (2019). Improving quality of care and outcome at very preterm birth: The Preterm Birth research programme, including the Cord pilot RCT. *Programme Grants for Applied Research*, *7*(8), 1–280
7	Mathematics	Mathematics students mathematical problem PISA learning student problems proof education solving reasoning math study activity knowledge thinking classroom maths number	1.40	1.59	Miyazaki, M., Fujita, T., & Jones, K. (2015). Flow-chart proofs with open problems as scaffolds for learning about geometrical proofs. *ZDM*, *47*, 1211–1224

TABLE 2 (Continued)

Topic number	Topic name	Characteristic words	REF2021 mean %	REF2014 mean %	Paper with highest proportion of words from the topic
8	Gender	Gender women girls boys male female men identity class women' sexual gendered feminist identities sex young masculinity university white performances	1.48	1.53	Ringrose, J., Tolman, D., & Ragonese, M. (2019). Hot right now: Diverse girls navigating technologies of racialized sexy femininity. *Feminism & Psychology*, *29*(1), 76–95
9	Regional issues and international development	Education international countries development global world Africa country south China comparative rural educational economic national UNESCO Hong African policy community	2.02	1.63	Chankseliani, M. (2021). The politics of exporting higher education: Russian university branch campuses in the "Near Abroad." *Post-Soviet Affairs*, *37*(1), 26–44
10	Psychiatry and Psychopathology	Emotional social problems study psychology mental model child development age psychological journal behaviour positive stress depression health effects time scores	2.41	1.35	Oliver, B. R., & Pike, A. (2018). Mother–child positivity and negativity: Family-wide and child-specific main effects and interactions predict child adjustment. *Developmental Psychology*, *54*(4), 744–756
11	Non-English Components	School education schools del prison national history Italian und Italy des German teaching social della educational der les literature educacion	0.40	0.45	Thyssen, G. (2015). Engineered communities? Industry, open-air schools, and imaginaries of belonging (c. 1913–1963). *History of Education & Children's Literature*, *10*(2), 297–320
12	Language and Linguistics	Language English learners languages linguistic learning Chinese words study university speakers word linguistics bilingual multilingual speech academic vocabulary acquisition native	1.86	1.60	Smith, E. (2016). Contact-induced change in a highly endangered language of northern Bougainville. *Australian Journal of Linguistics*, *36*(3), 369–405
13	Technology Enhanced Learning	Learning technology digital online data media technologies open design mobile learners social educational information activities computer engagement internet support project	2.70	2.65	Ferguson, R., & Clow, D. (2015). Consistent commitment: Patterns of engagement across time in Massive Open Online Courses (MOOCs). *Journal of Learning Analytics*, *2*(3), 55–80
14	Critical and Social Theory	Social education cultural research knowledge educational practices identity journal people capital power critical university sociology London theory young ways studies	5.82	5.77	Costa, C., Burke, C., & Murphy, M. (2019). Capturing habitus: Theory, method and reflexivity. *International Journal of Research & Method in Education*, *42*(1), 19–32
15	History, Religion and Race	Education religious history black race white religion British London ethnic university racism schools racial historical Christian England faith Britain church	1.75	2.11	Bartie, A., Fleming, L., Freeman, M., Hulme, T., Hutton, A., & Readman, P. (2019). 'History taught in the pageant way': Education and historical performance in twentieth-century Britain. *History of Education*, *48*(2), 156–179

(Continues)

TABLE 2 (Continued)

Topic number	Topic name	Characteristic words	REF2021 mean %	REF2014 mean %	Paper with highest proportion of words from the topic
16	Early Childhood and Families	Children children' child parents early childhood family years families development young play home education school parental research parent child' rights	2.83	2.77	Katsiada, E., Roufidou, I., Wainwright, J., & Angeli, V. (2018). Young children's agency: Exploring children's interactions with practitioners and ancillary staff members in Greek early childhood education and care settings. *Early Child Development and Care*, *188*(7), 937–950
17	Leadership and Management	Leadership management development leaders professional change network work practice role educational social learning networks community staff partnership collaboration school support	2.17	2.43	Daly, A. J., Liou, Y. H., & Brown, C. (2016). Social red bull: Exploring energy relationships in a school district leadership team. *Harvard Educational Review*, *86*(3), 412–448
18	New Materialism	Space research arts art body time visual ways practices spaces work place creative material world studies London university movement images	3.51	3.01	Hackett, A., & Somerville, M. (2017). Posthuman literacies: Young children moving in time, place and more-than-human worlds. *Journal of Early Childhood Literacy*, *17*(3), 374–391
19	Clinical Psychology and Developmental Disorders	Autism children ASD disorders developmental social child autistic spectrum disorder studies journal group study difficulties diagnosis ADHD disabilities research intellectual	1.33	0.84	Christiansz, J. A., Grey, K. M., Taffe, J., & Tonge, B. J. (2016). Autism spectrum disorder in the DSM-5: Diagnostic sensitivity and specificity in early childhood. *Journal of Autism and Developmental Disorders*, *46*, 2054–2063
20	Claims of Significance	First significant different findings specific identified influence field journal effects find staff effect differences significantly five reflect confidence doi specifically	1.57	0.94	Baird, J. A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 44–59
21	Sports	Physical activity sport coaching health outdoor food activities environmental environment study literacy sports time participants bmi doi exercise movement coach	1.04	0.56	Evans, K. L., Hughes, J., & Williams, M. D. (2018). Reduced severity of lumbo-pelvic-hip injuries in professional Rugby Union players following tailored preventative programmes. *Journal of Science and Medicine in Sport*, *21*(3), 274–279
22	Analysing Large-Scale Data	Data model age table variables results effects social effect level analysis differences sample models income average educational higher cohort class	4.85	3.60	Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modelling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioural Statistics*, *39*(5), 307–332
23	Schooling Systems	School schools pupils students educational teachers education secondary primary year attainment England pupil research achievement teacher children British English curriculum	4.45	5.16	See, B. H., Morris, R., Gorard, S., Kokotsaki, D., & Abdi, S. (2020). Teacher recruitment and retention: A critical review of international evidence of most promising interventions. *Education Sciences*, *10*(10), 262

TABLE 2 (Continued)

Topic number	Topic name	Characteristic words	REF2021 mean %	REF2014 mean %	Paper with highest proportion of words from the topic
24	Affective Factors	Students learning motivation achievement academic items psychology feedback study performance educational model journal doi positive self-efficacy engagement beliefs factor anxiety	2.43	1.73	Putwain, D. W., Remedios, R., & Symes, W. (2015). Experiencing fear appeals as a challenge or a threat influences attainment value and academic self-efficacy. *Learning and Instruction*, *40*, 21–28
25	Children's Social Care	Young health people social care youth children mental services poverty support child family life well-being work wellbeing risk violence abuse	2.23	1.93	McCartan, C., Bunting, L., Bywaters, P., Davidson, G., Elliott, M., & Hooper, J. (2018). A four-nation comparison of kinship care in the UK: The relationship between formal kinship care and deprivation. *Social Policy and Society*, *17*(4), 619–635
26	Education Policy	Education policy government public national state London local university system policies political England educational governance economic provision sector funding standards	4.72	5.33	Hall, D., Grimaldi, E., Gunter, H. M., Møller, J., Serpieri, R., & Skedsmo, G. (2015). Educational reform and modernisation in Europe: The role of national contexts in mediating the new public management. *European Educational Research Journal*, *14*(6), 487–507
27	Higher Education	Students higher education university student academic universities research study institutions teaching academics participation institutional degree learning studies access undergraduate institution	3.65	3.78	Elliott, G. (2019). Widening participation, student identity and agentic capital in coastal, rural and isolated communities in south-west England. *Widening Participation and Lifelong Learning*, *21*(1), 117–138
28	Training and Employment	Work education labour skills social employment market job training graduates vocational economic career working countries jobs higher time adult workers	2.37	2.23	Green, F., Felstead, A., Gallie, D., Inanc, H., & Jewson, N. (2016). The declining volume of workers' training in Britain. *British Journal of Industrial Relations*, *54*(2), 422–448
29	Special Educational Needs and Disabilities	Disability inclusive special education inclusion disabilities children music disabled support educational people journal mainstream social send sen learning musical attitudes	1.17	0.96	Ravenscroft, J., Davis, J., Bilgin, M., & Wazni, K. (2019). Factors that influence elementary school teachers' attitudes towards inclusion of visually impaired children in Turkey. *Disability & Society*, *34*(4), 629–656
30	Teacher Education and Professional Development	Teachers learning teacher teaching education practice professional knowledge curriculum development students classroom student research pedagogy journal skills university practices learners	5.90	7.33	Perry, E., & Boylan, M. (2018). Developing the developers: Supporting and researching the learning of professional development facilitators. *Professional Development in Education*, *44*(2), 254–271

(Continues)

TABLE 2 (Continued)

Topic number	Topic name	Characteristic words	REF2021 mean %	REF2014 mean %	Paper with highest proportion of words from the topic
31	Science Education	Science students scientific stem education physics scientists game creativity engineering journal inquiry knowledge capital study games e.g. learning teaching understanding	1.25	1.19	To, C., Tenenbaum, H. R., & Hogh, H. (2017). Secondary school students' reasoning about evolution. *Journal of Research in Science Teaching*, *54*(2), 247–273
32	Interviews and Focus Groups	Research participants data experiences work group study experience time interviews people interview it' qualitative focus felt support personal part feel	9.34	11.11	Willis, J., & Baines, E. (2018). The perceived benefits and difficulties in introducing and maintaining supervision groups in a SEMH special school. *Educational Review*, *70*(3), 259–279
33	Systematic Reviews and Meta-Analyses	Intervention interventions risk review studies health control outcomes study bias women smoking group outcome effect data Cochrane published systematic low	1.13	0.90	Chamberlain, C., O'Mara-Eves, A., Porter, J., Coleman, T., Perlen, S. M., Thomas, J., & McKenzie, J. E. (2017). Psychosocial interventions for supporting women to stop smoking in pregnancy. *Cochrane Database of Systematic Reviews*, *2*(2), CD001055
34	Cognitive Processing	Task memory learning participants cognitive group effects processing performance effect control training tasks attention working time condition doi brain study	2.04	1.57	Longman, C. S., Lavric, A., & Monsell, S. (2017). Self-paced preparation for a task switch eliminates attentional inertia but not the performance switch cost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 862
35	Reading	Reading language literacy children comprehension skills word vocabulary words test read writing time group text knowledge phonological awareness development scores	1.78	1.93	Hulme, C., Nash, H. M., Gooch, D., Lervåg, A., & Snowling, M. J. (2015). The foundations of literacy development in children at familial risk of dyslexia. *Psychological Science*, *26*(12), 1877–1886

papers which had 0% of words from Topic 20, we found instances of papers which made little attempt to draw wider implications. For example, Langdown et al.'s (2019) article 'Acute effects of different warm-up protocols on highly skilled golfers' drive performance' provided compelling evidence for how golfers might improve their drives, but did not attempt to generalise to learning sporting skills outside of golf, or to learning more generally. We decided to name Topic 20 Claims of Significance.

A full dataset showing the topic compositions for the 4290 English-language journal articles submitted to REF2021 is available online at https://doi.org/10.17028/rd.lboro.25201139.v1. Studying these data alongside Table 2 will allow readers to investigate the extent to which they feel our topic names appropriately capture the meaning of each topic.[i]

The next step was to calculate each submitted unit's mean proportions for each topic. This gave us a representation of the overall proportion of each unit's submission from each topic area. For instance, 17.4% of the 'composite mean paper' returned by the Open University (an imagined paper composed of the same topic weightings as the mean topic weightings of the actual papers returned by the Open University) was made up of words from the Technology Enhanced Learning topic. Similarly, 18.9% of the composite mean paper returned by Manchester Metropolitan University was made up of words from the New Materialism topic, and 14.7% of Loughborough University's composite mean paper was from the Mathematics topic. These results, and other comparable figures, seemed consistent with our impressions of the research strengths of these departments, providing some evidence of the face validity of our model. The mean topic weightings, across all topics, for each institution submitted to the education subpanel—together with their output quality profiles, output GPA, number of FTE staff submitted and proportion of eligible staff submitted—are available online at https://doi.org/10.17028/rd.lboro.25201139.v1.

Table 2 shows the overall mean proportion (across all 4290 submitted papers to REF2021) for each topic. These figures give an overall sense of the makeup of education research papers submitted to REF2021. They show that Interviews and Focus Groups was the most popular topic (9.34%), followed by Methodological Depth (7.99%), Teacher Education and Professional Development (5.90%), Critical and Social Theory (5.82%) and Philosophy of Education (5.75%).

Next, we evaluated the extent to which our model (i.e., 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 these topic proportions sum to 1 for each unit, we could not run a standard regression (there would have been perfect multicollinearity). Instead, we adopted a compositional regression approach using base 2 additive log-ratios (Coenders & Pawlowsky-Glahn, 2020). Because we analysed every English-language journal article returned to REF2021, we conceptualise this regression as being a whole-population analysis (Berk, 2004, p. 42), the goal of which is to understand judgements made in REF2021. We therefore do not report inferential statistics. This conceptualisation raises the issue of whether the resulting model can accurately predict independent judgements of research quality (i.e., those made outside the context of REF2021). We consider this issue later in the paper by using our model to analyse REF2014 submissions.

Our regression analysis yielded two results of interest. First, we calculated the overall fit of the compositional regression model, which can be interpreted as telling us how much of the variance in output GPAs can be collectively explained by the 35 topics. We also ran a model where the proportion of 4* outputs was the dependent variable, which yielded essentially identical results. Our model explained a large proportion of the variance in units' output GPAs, $R^2 = 84.1\%$. In other words, knowing the content of the composite mean journal article returned by each unit allowed us to predict with a very high degree of accuracy the output GPA they received from the expert peer review process. This is particularly striking given that journal articles made up only 81.5% of the outputs that were submitted to the subpanel.

However, given the limited number of units (83) and the large number of topics needed to characterise the diversity of education research (35), it is possible that this very large R^2 is the result of overfitting. In other words, despite this very large R^2, it is possible that, if we applied our model to a new sample of units, then it would not accurately predict quality judgements of their outputs. We return to this issue later in the paper by applying our model to REF2014 to evaluate the extent to which it can predict out-of-sample judgements of quality.

Second, we calculated the regression coefficients associated with each of the 35 topics. In a base 2 additive log-ratio compositional regression, the coefficient associated with each predictor is interpreted as the expected change in the dependent variable (output GPA in our case) if the value of the predictor doubles, with the remaining predictors all reducing proportionately (i.e., retaining their relative ratios). For instance, the regression coefficient associated with the Gender topic was 0.035. This means that if one unit's composite mean paper had twice as much content about gender as another unit's, with both having an identical balance of the remaining topics, our model would predict that the first unit's output GPA would be 0.035 higher than the second's. The full regression model is shown in Table 3.

The regression coefficients varied from +0.104 (Analysing Large-Scale Data) to −0.164 (Teacher Education and Professional Development). We interpreted the effect of a topic to be 'large' if it had an absolute coefficient above 0.05 (a gap equivalent to the difference between the output GPAs obtained by the University of Sussex (ranked 14th by output GPA) and the University of Southampton (ranked 18th)). Five topics had large positive coefficients: Analysing Large-Scale Data (0.104), Methodological Depth (0.104), Critical and Social Theory (0.092), Claims of Significance (0.081) and Schooling Systems (0.068). Three topics had large negative coefficients: Interviews and Focus Groups (−0.077), Higher Education (−0.102) and Teacher Education and Professional Development (−0.164). In other words, returns that contained many words associated with the analysis of large-scale data, with significant methodological discussion, with ideas from critical or social theory, or with claims of originality and significance, typically received higher scores than those which did not. Conversely, returns that included more words associated with interviews or focus groups, or with higher education or teacher education/PD, on average received lower scores than those which included fewer. These findings are difficult to reconcile with the REF subpanel's claim that there was 'no strong association between research excellence and particular methods or approaches' (REF, 2022, p. 159). We return to this issue in the discussion.

In sum, using our model we were able to successfully explain a surprisingly large proportion of the variance in units' output GPAs. This allowed us to draw two main conclusions. First, we could see the popularity of topics, methods and approaches used by educational researchers in the United Kingdom, at least within the subset of journal articles selected to be returned to REF2021. Second, we were able to identify those topics, methods and approaches that were associated with judgements of higher quality made by the REF2021 education subpanel (and those that were associated with judgements of lower quality).

To address three remaining research issues—relating to (i) documenting changes in research focus over time, (ii) assessing whether our REF2021 model's high R^2 was down to overfitting and (iii) evaluating the level of between-panel stability in quality judgements—we applied our model to the previous REF exercise (REF2014).

APPLYING THE MODEL TO REF2014

Compared to 2021, slightly fewer universities made returns to the education subpanel in 2014 (75 compared to 83). However, these 75 units submitted more outputs—a total of 5519. Of these, 4322 (78.3%) were self-declared to be journal articles, of which 12 were written in a language other than English and excluded from our analysis. This left 4310 journal articles,

TABLE 3 A compositional regression predicting REF2021 output GPAs with our 35 topics.

Predictor	Regression coefficient
(Intercept)	2.953
Analysing Large-Scale Data	0.104
Methodological Depth	0.104
Critical and Social Theory	0.092
Claims of Significance	0.081
Schooling Systems	0.068
Communication and Interaction	0.047
Gender	0.035
History, Religion and Race	0.029
Reading	0.024
Non-English Components	0.020
Psychiatry and Psychopathology	0.019
Mathematics	0.010
Science Education	0.003
Special Educational Needs and Disabilities	0.003
Children's Social Care	−0.002
New Materialism	−0.004
Philosophy of Education	−0.004
Cognitive Processing	−0.004
Health and Medicine	−0.004
Sports	−0.008
Citizenship and Culture	−0.009
Training and Employment	−0.014
Developmental Psychology	−0.015
Systematic Reviews and Meta-Analyses	−0.018
Education Policy	−0.018
Language and Linguistics	−0.018
Affective Factors	−0.021
Clinical Psychology and Developmental Disorders	−0.024
Technology Enhanced Learning	−0.024
Regional Issues and International Development	−0.029
Leadership and Management	−0.031
Early Childhood and Families	−0.050
Interviews and Focus Groups	−0.077
Higher Education	−0.102
Teacher Education and Professional Development	−0.164
	$R^2 = 0.841$

Note: Topics are ordered by the size of the regression coefficient.

of which we were able to obtain pdf copies of 4269 (99.0%, a lower proportion than the 100% figure we achieved for REF2021, perhaps because of the more stringent open access rules that were introduced for the latter exercise). We converted these 4269 articles into plain text using the UNIX pdftotext command (Poppler, 2022), and used our 35-topic model derived from the REF2021 papers to calculate the composition of each article. A full dataset showing the topic compositions for the 4269 articles in our REF2014 sample is available online at https://doi.org/10.17028/rd.lboro.25201139.v1.

We asked three main questions. First, what changes can be observed in the frequencies with which the various topics were represented in papers submitted to the 2014 and 2021 REFs? Did certain topics become more or less prominent over this period? Second, can the model trained on 2021 papers successfully predict the unit-level output GPAs achieved by the 2014 papers, as assigned by the REF2014 education subpanel? If the answer to this latter question is yes, then we should have confidence that our model is able to predict out-of-sample judgements of research quality (i.e., the large R^2 observed in the context of REF2021 cannot solely be due to overfitting) and also that the judgements made by the 2014 panel are similar to those made by the 2021 panel, despite the substantial changes in panel membership.

The mean proportion of words (averaged across all articles in REF2014) from each topic is shown in the fifth column of Table 2. There were notable changes between 2014 and 2021. In general, British education research, at least as represented by the papers chosen to be returned to the REF, seems to have become more quantitative and less qualitative between 2014 and 2021. Specifically, over this period the Interviews and Focus Groups topic declined in prominence by 16% (11.11% to 9.34%), and there were increases in the prominence of the Analysing Large-Scale Data (3.60% to 4.85%, an increase of 35%) and Systematic Reviews and Meta-Analyses (0.90% to 1.13%, an increase of 26%) topics, as well as increases in the prominence of psychological topics that one might expect to be associated with quantitative methods. For instance, the Psychiatry and Psychopathology, Clinical Psychology and Developmental Disorders, Cognitive Processing and Developmental Psychology topics all saw increases of over 10%. In terms of curriculum areas, compared to 2014, 2021 saw an increase in the quantity of research on Sports (0.56% to 1.04%, an increase of 88%), Citizenship and Culture (1.14% to 1.40%, an increase of 22%) and Language and Linguistics (1.60% to 1.86%, an increase of 16%), but decreases in History, Religion and Race (2.11% to 1.75%, a decrease of 17%) and Mathematics (1.59% to 1.40%, a decrease of 12%).

Next, we calculated the mean composite paper associated with each of the 75 submissions made to the REF2014 education subpanel in a similar manner to our REF2021 analysis, by taking the mean composition for each topic across all papers submitted by each unit. We then used the regression coefficients shown in Table 3 (from our model predicting output GPAs in 2021) to calculate predicted output GPAs. This allowed us to produce 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.

We then compared these predicted output GPAs with the actual output GPAs assigned by the REF2014 education subpanel. The predicted and actual output GPAs for each of the units that submitted to the 2014 education subpanel are shown in Figure 3. The correlation between the predicted and actual output GPAs was high, at $r=0.658$, $R^2=43.3\%$. We make two remarks. First, given the difference between the rules on output selection used in REF2014 and REF2021 (in REF2014 every researcher returned was required to submit four articles, in REF2021 this could vary between one and five), we might expect this correlation to be weaker than if the same rules had been adopted for both exercises. Second, our model was considerably better at predicting 2014 output GPAs than is typically achieved by citation analyses. For instance, Pride and Knoth (2018) used median concurrent citation counts to predict units' output GPAs in REF2014. They found a correlation of $r=0.469$, $R^2=22.0\%$ between the median number of citations achieved by units' submitted papers (as of 2014) and their output GPAs in the education subpanel. In other words, our topic model explained

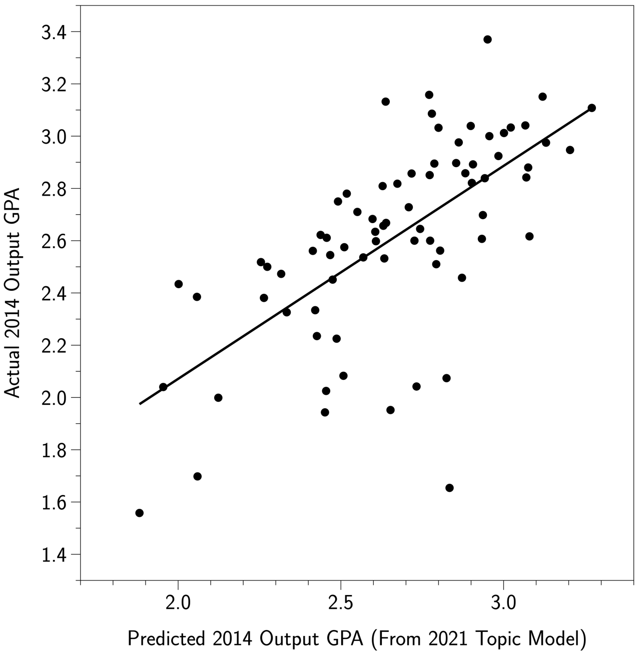

FIGURE 3 A plot of predicted REF2014 output GPAs from our 2021 topic model, against the output GPAs assigned by the REF2014 subpanel. GPA, grade point average; REF, Research Excellence Framework.

around twice as much variance in REF2014 output GPAs as the citation methods used by Pride and Knoth. It is worth highlighting that, because our main aim was to interpret the nature of the relationships between topic use and output GPAs, we used a relatively simple linear regression. If our goal was simply to produce a model with the highest R^2 possible, we could have adopted a more complex, but also more opaque, modelling approach.

In sum, by analysing the journal articles submitted to REF2014 using the topic model derived from the journal articles submitted to REF2021, and applying the same regression coefficients to predict units' output GPAs, we were able to produce a reasonably accurate estimate of how the articles submitted to REF2014 were assessed by the 2014 panel, suggesting that the large R^2 observed in our REF2021 model was not solely due to overfitting, and also that there was a reasonable degree of consistency in the approaches used by the two subpanels to assess research quality.

DISCUSSION

Summary of main findings

In order to gain insights into education research in the United Kingdom, and its perceived quality, we analysed the full texts of all journal articles submitted to the education subpanel of REF2021, the high-stakes research quality assessment exercise conducted by government research funding agencies. Using latent Dirichlet allocation topic modelling, we identified 35 topics that together provide a summary of the issues focused upon, and the approaches adopted, by UK-based education researchers, or at least by that subsection of outputs chosen for submission to the REF. By analysing the composition of each submitted journal article in terms of these 35 topics, we established four main findings.

First, the semantic content of the journal articles that a unit decided to submit to the REF was predictive of the quality assessment scores—designed to capture the originality, significance and rigour of the submitted outputs—that the unit received from the expert peer review process. Specifically, our model explained 84.1% of the variance in unit-level output GPAs in REF2021. This is particularly notable given that we analysed only journal articles, which comprised just 81.5% of the outputs submitted to the education subpanel.

Second, we were able to establish which of the 35 topics were particularly strongly predictive of quality judgements made by the panel. From this analysis we concluded that returns which included many papers that analysed large-scale data, that had detailed critical discussions of methodological issues, that adopted critical or social theories, that analysed schooling systems or that made strong arguments for their originality or significance, on average received higher scores than returns which included fewer papers with these features. Similarly, returns which include more papers that adopted interviews and focus groups, that focused on higher education or that analysed teacher education and professional development on average received lower scores than those which included fewer. Notably, these findings seem to conflict with the subpanel's claim that there was ‘no strong association between research excellence and particular methods or approaches’ (REF, 2022, p. 159). We found several such associations.

Third, by applying our model to the full text of 99% of journal articles submitted to the education subpanel of the previous assessment exercise (REF2014), we were able to identify topics which have increased in prominence, and topics which have decreased. We found evidence of a shift towards quantitative methods and away from qualitative methods. But, despite this shift, qualitative methods are still extremely common in UK-based education research: the Interviews and Focus Groups topic remains the most widely used of all 35 topics.

Finally, to assess whether our model could predict independent judgements of research quality, and also to evaluate the extent to which the high R^2 we observed for our 2021 model was due to overfitting, we attempted to predict the unit-level output GPAs we would have expected each unit to have received in REF2014, based on the composition of the papers they submitted to that exercise, using only our 2021 model. Our predicted output GPAs were strongly correlated with the output GPAs assigned by the 2014 subpanel, suggesting (i) that the two panels made their judgements based on largely similar criteria and (ii) that topic modelling is able to successfully predict the research quality judgements of collections of manuscripts made in a separate exercise.

Quality differences between different research approaches

We found strong associations between the extent to which a unit submitted papers that adopted particular research methods or approaches, and the scores they received. Interview- and focus group-based research was associated with lower scores, and large-scale data analyses, systematic reviews and meta-analyses were associated with higher scores. However, our analysis demonstrates associations, not directional causal relationships.

These findings need unpacking. In particular, it would be tempting to interpret these results in terms of a quantitative/qualitative hierarchy. However, although the Interviews and Focus Groups topic clearly comprises qualitative research, so did other topics. The Communication and Interaction topic, for instance, contained other kinds of qualitative research, such as text, discourse and interaction analyses, conducted on other kinds of datasets, such as transcripts of classroom interactions. Unlike with the Interviews and Focus Groups topic, units that focused more on communication and interaction tended to receive higher scores (i.e., this topic had a positive regression coefficient in our model).

In sum, our analysis showed that it is not qualitative research per se that the panel, on average, gave lower scores to, but rather qualitative analyses of interview and focus group data.

Nevertheless, given the history of the so-called paradigm wars in education research (e.g., Galvez et al., 2019), and controversies about perceived hierarchies of research methods (e.g., Ercikan & Roth, 2006; Tooley & Darby, 1998), the finding that units which return more interview- or focus group-based outputs appear to receive systematically lower scores is particularly notable. Figure 4 shows the relationship between units' mean proportions of words from the Interviews and Focus Groups topic and their output GPAs in both REF2021 and REF2014. Although these relationships are extremely strong, $r = -0.585$ and -0.399, most high-performing units returned some interview-based papers. Of the top nine units by output GPA (who each received a GPA of 3.29 or above), only the University of Durham returned no heavily interview-based outputs (operationalised here as articles with more than 20% of their words from the Interviews and Focus Groups topic). The other eight collectively returned 44 such outputs (8% of their total), providing some suggestive evidence that it is certainly possible for interview studies to receive high scores.

Although the correlation between the Interviews and Focus Groups topic proportion and output GPA was strong, it is possible that this relationship was driven by factors that covary with both. For instance, one explanation for the relationship shown in Figure 4 is that interviews and focus groups are typically used more frequently in generally weaker research domains. Indeed, the two topics with the strongest positive correlations with the Interviews and Focus Groups topic were Higher Education ($r = 0.348$) and Teacher Education and Professional Development ($r = 0.405$), which were themselves the two topics that had the largest negative regression coefficients in our analysis. So, perhaps the reason interviews and focus groups seem to have been judged negatively by the panel is that they were disproportionately often used in weak research domains, and it is these research domains

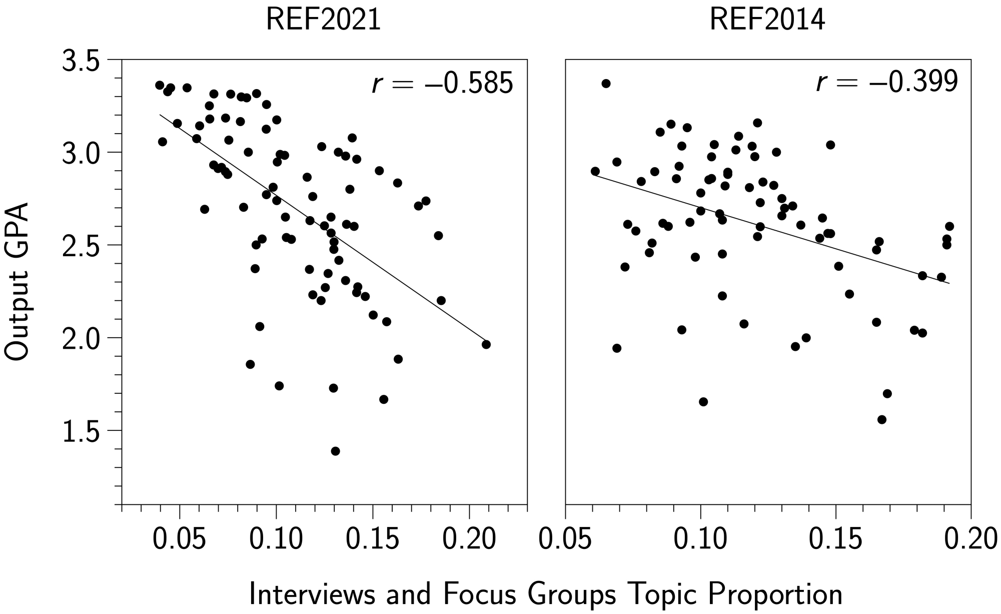

FIGURE 4 A plot of the mean proportion of words from the Interviews and Focus Groups topic that units returned in REF2021 and REF2014 against their output GPAs, together with the lines of best fit. GPA, grade point average; REF, Research Excellence Framework.

that drive quality judgements rather than the use of interviews or focus groups. However, conversely, we also cannot rule out the possibility that these domains were deemed weak because they involved more of these methods. Without access to individual output scores, or the ability to run an experimental study to establish causality, it is difficult to disentangle these possibilities further. But notably, our findings are consistent with results from researchers who were, as part of a Research England project, given access to the confidential judgements on individual outputs from across the entire REF. Thelwall et al. (2023a) analysed the titles and abstracts (but not the full texts) of journal articles submitted to all disciplinary subpanels (not just education) in REF2021, finding that papers which included words associated with qualitative research in their abstracts typically received lower quality judgements than those that did not.

Although we cannot confidently establish the mechanism behind the relationship between a unit's use of interviews or focus groups and its output GPA, we have robustly demonstrated the existence of this association. Given this, why did the subpanel assert that no strong associations between approach and quality were present in their assessments? One straightforward possibility is that the subpanel was simply not aware of these associations: detecting them 'by eye' might well be extremely difficult. Perhaps, for example, these relationships were disguised by the presence of particularly salient counterexamples (interview-based papers that received 4* assessments and systematic reviews that received 1* assessments). One advantage of the topic modelling approach is that it reveals relationships that may be difficult to detect through other methods. Another possibility is that the panel were in fact aware of these associations but wanted to avoid strongly influencing submissions to future assessment exercises by drawing attention to them.

Changes over time

We found evidence that the content of submissions to the education subpanel has changed systematically between the 2008–2014 period and the 2015–2021 period. Some of these changes are likely to reflect top-down initiatives. For instance, we found an increase in the number of outputs that analysed large-scale data. This approach to research has been strongly encouraged over the last 15 years by the Economic and Social Research Council (ESRC), the main responsive-mode education funding body in the United Kingdom. In 2011, the ESRC launched its Secondary Data Analysis Initiative to 'create opportunities for researchers to exploit existing national datasets' (ESRC, 2012, p. 26). This led to a series of regular funding calls that were dedicated to funding research that analysed large-scale secondary data. Similarly, we found an increase in the prominence of systematic reviews and meta-analyses. This is likely to reflect the influence of the Pupil Premium Toolkit, launched by the Education Endowment Foundation and Sutton Trust in 2011 (Higgins et al., 2011). Since then, the Toolkit has had a remarkable impact on policy and practice: it is used by two-thirds of headteachers in the United Kingdom and is regularly cited in government policy documents (University of Durham, 2022). Given this, it is perhaps unsurprising that the wider field has seen increased interest in the use of systematic reviews and meta-analyses.

Other changes between 2014 and 2021 seem harder to explain. For instance, given successive governments' emphasis on the importance of improving mathematics education in the United Kingdom (e.g., Industrial Strategy, 2018), it is surprising that 12% less mathematics education research was submitted to REF2021 than REF2014. Similarly, the reasons behind the substantial increase in the amount of education research focused on sports are not clear.

Should topic modelling be used to predict REF scores?

We were able to use our REF2021 topic model to predict the output quality scores assigned to REF2014 submissions. This suggests that there is some degree of between-REF consensus about the construct of research quality. This finding is particularly notable given that there was relatively little overlap between the membership of the 2014 and 2021 education subpanels: of the 36 REF2021 subpanel members, only 4 (11%) had served on the REF2014 subpanel. In light of academics' commitment to peer review as the best way of assessing research quality (e.g., Rowley & Sbaffi, 2018), and given that the correspondence between our model's predictions and the actual REF2014 was high but far from perfect (cf. Thelwall et al., 2023b), it seems unlikely that statistical analyses of the sort that we have conducted here could replace peer review in future REFs. Nevertheless, the fact that our model was apparently able to successfully give insights into quality judgements made during a different exercise by a largely different group of reviewers raises the prospect of using our model to assist with the preparation of future REF submissions. One could use our model to generate predicted REF scores for candidate outputs, and simply return those with the highest predictions (e.g., our model predicts that the current paper would receive a rating of 3.27 in a future REF exercise). But would it be sensible to use the model in this way?

One difficulty with this proposal is that, by necessity given the confidential nature of REF scores, we were only able to predict unit-level output GPAs, not output-level quality judgements. In other words, we used ecological correlations: the correlation between two variables that are themselves group means (in our case, unit-level output GPAs and the topic weightings of units' composite mean papers). Ecological correlations are often stronger than the equivalent correlations calculated on individual data (e.g., Hammond, 1973; Robinson, 2009), and assuming that these two correlations are the same is a mistake known as the ecological fallacy. The fallacy can be illustrated 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 citation counts and REF2014 quality judgements reported by Wilsdon et al. (2015) in their REF-commissioned study of whether metrics could replace expert peer review in the REF. For the education subpanel, Pride and Knoth reported a correlation of 0.414 between units' mean citation counts and their output GPAs, whereas Wilsdon et al. found an individual-level correlation of 0.183. An analogous reduction of the 0.658 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 we cannot estimate the size of the reduction with any accuracy.

Given the ecological fallacy, we doubt that drawing strong conclusions about individual outputs on the basis of a model like ours can be justified. Nevertheless, the REF is an assessment of research groups, not of individual research outputs. It is the unit-level scores that are published and which influence future research funding. Arguably then, it is the group level, not the output level, that analysts should focus upon. Given this, it is important that we have demonstrated that our model accurately predicts performance at the group level (and substantially more so than citation analyses, which we know that at least some universities did use to inform their REF submission strategies; Manville et al., 2021).

A more defensible approach to using topic modelling in the preparation of REF submissions might be to use a model of the kind we have presented here to assess whether a unit's internal selection process for a future REF is generating implausibly high or low scores. For instance, if a unit decided to use peer review to select outputs for REF2029, and if those internal scores led to a predicted output GPA dramatically different from that predicted by our topic model, then this might be reason to conduct some additional calibration checks, perhaps using additional external reviewers. Certainly, our findings suggest that using a topic model for this kind of secondary checking purpose is likely to be more useful than relying on citation metrics.

ACKNOWLEDGEMENTS

We are grateful to Barbara Jaworski, Charles Crook, Sam Sims and two anonymous reviewers for insightful comments on earlier drafts of this paper.

FUNDING INFORMATION

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].

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Online materials associated with this manuscript are available at https://doi.org/10.17028/rd.lboro.25201139.v1.

ETHICS STATEMENT

There are no ethical issues raised.

ORCID

Matthew Inglis https://orcid.org/0000-0001-7617-4689

ENDNOTE

[i] Note that where a particular paper was returned by one or more units (where it had coauthors from multiple institutions), we treated these instances independently. So, there are duplicate instances of certain papers in our dataset. Because MALLET uses Gibbs sampling (a stochastic process) when fitting topics to data, these duplicate instances of the same papers should be expected to have very slightly different topic proportions (perhaps in the third or fourth decimal place).

REFERENCES

Baird, J. A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 44–59.

Bennett, A., Inglis, M., & Gilmore, C. (2019). The cost of multiple representations: Learning number symbols with abstract and concrete representations. *Journal of Educational Psychology*, *111*(5), 847–860.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Sage.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Brassington, L. (2022). *Research evaluation: Past, present and future*. HEPI report 152. Higher Education Policy Institute.

Brown, R., & Carasso, H. (2013). *Everything for sale? The marketisation of UK higher education*. Routledge.

Coenders, G., & Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT*, *44*, 201–220. https://doi.org/10.2436/20.8080.02.100

Ercikan, K., & Roth, W. M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23.

ESRC. (2012). *Shaping society: Economic and Social Research Council annual report and accounts 2011/12*. https://assets.publishing.service.gov.uk/media/5a7ca88d40f0b6629523afb0/0338.pdf

Fairclough, N. (1995). *Critical discourse analysis*. Longman.

Galvez, S., Heiberger, R., & McFarland, D. (2019). Paradigm wars revisited: A cartography of graduate research in the field of education (1980–2010). *American Educational Research Journal*, *57*(2), 612–652. https://doi.org/10.3102/0002831219860511

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

Gibbs, S., & Elliott, J. (2015). The differential effects of labelling: How do 'dyslexia' and 'reading difficulties' affect teachers' beliefs. *European Journal of Special Needs Education*, *30*(3), 323–337.

Gillies, D. (2008). *How should research be organised?* College Publications.

Hammond, J. L. (1973). Two sources of error in ecological correlations. *American Sociological Review*, *38*, 764–777.

Higgins, S., Kokotsaki, D., & Coe, R. (2011). *Toolkit of strategies to improve learning: Summary for schools spending the pupil premium and technical appendices*. Sutton Trust.

Hitchcock, J. H., & Onwuegbuzie, A. J. (2020). Developing mixed methods crossover analysis approaches. *Journal of Mixed Methods Research*, *14*(1), 63–83.

Industrial Strategy. (2018). *Industrial strategy: Building a Britain fit for the future*. https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future

Inglis, M., & Foster, C. (2018). Five decades of mathematics education research. *Journal for Research in Mathematics Education*, *49*(4), 462–500.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*, 89–106.

Langdown, B. L., Wells, J. E., Graham, S., & Bridge, M. W. (2019). Acute effects of different warm-up protocols on highly skilled golfers' drive performance. *Journal of Sports Sciences*, *37*(6), 656–664.

Manville, C., d'Angelo, C., Culora, A., Gloinson, E. R., Stevenson, C., Weinstein, N., et al. (2021). *Understanding perceptions of the Research Excellence Framework among UK researchers: The real-time REF review*. RAND Corporation.

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.

McCallum, A. K. (2002). *MALLET: Machine learning for language toolkit*. http://mallet.cs.umass.edu

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education*, *31*(4), 297–311.

Myhill, D., & Jones, S. (2015). Conceptualizing metalinguistic understanding in writing/conceptualización de la competencia metalingüística en la escritura. *Cultura y Educación*, *27*(4), 839–867.

Nelson, J. (2017). Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qualitative Research*, *17*(5), 554–570.

Pardo-Guerra, J. P. (2022). *The quantified scholar: How research evaluations transformed the British social sciences*. Columbia University Press.

Pinar, M., & Horne, T. J. (2022). Assessing research excellence: Evaluating the Research Excellence Framework. *Research Evaluation*, *31*(2), 173–187.

Poppler. (2022). *Computer software*. https://poppler.freedesktop.org

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. Proceedings of the 22nd International Conference on Theory and Practice in Digital Libraries, TPDL 2018, Porto, Portugal, September 10–13*. Springer. https://doi.org/10.1007/978-3-030-00066-0_17

REF. (2019). *REF panel criteria and working methods*. UK Higher Education Funding Bodies. https://www.ref.ac.uk/media/1450/ref-2019_02-panel-criteria-and-working-methods.pdf

REF. (2022). *Overview report by main panel C and sub-panels 13 to 24*. UK Higher Education Funding Bodies. https://www.ref.ac.uk/media/1912/mp-c-overview-report-final-updated-september-2022.pdf

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, *38*(2), 337–341.

Rowley, J., & Sbaffi, L. (2018). Academics' attitudes towards peer review in scholarly journals and the effect of role and discipline. *Journal of Information Science*, *44*(5), 644–657.

Sechelski, A. N., & Onwuegbuzie, A. J. (2019). A call for enhancing saturation at the qualitative data analysis stage via the use of multiple qualitative data analysis approaches. *The Qualitative Report*, *24*(4), 795–821.

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. M. (2023a). Terms in journal articles associating with high quality: Can qualitative research be world-leading? *Journal of Documentation*, *79*, 1110–1123.

Thelwall, M., Kousha, K., Wilson, P., Makita, M., Abdoli, M., Stuart, E., et al. (2023b). Predicting article quality scores with machine learning: The UK Research Excellence Framework. *Quantitative Science Studies*, *4*(2), 547–573.

Tooley, J., & Darby, D. (1998). *Educational research—A critique*. Office for Standards in Education.

University of Durham. (2022). *The Pupil Premium Toolkit: Evidence for impact in education*. REF Impact Case Study. https://results2021.ref.ac.uk/impact/efda03e1-b57a-417b-b97d-72cd0d99529d?page=1

Watermeyer, R. P., & Derrick, G. (2022). Affective auditing: The emotional weight of the Research Excellence Framework on middle management. *Research Evaluation*, *31*, 498–506. https://doi.org/10.1093/reseval/rvac041

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., et al. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. HEFCE.

How to cite this article: 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*, *00*, 1–24. https://doi.org/10.1002/berj.4040