ChatGPT Ranking of Business and Management Journals With Article Quality Scores[1]

Mike Thelwall, Information School, University of Sheffield, UK.

Purpose: Business and management journal rankings are controversial but influential for scholars seeking publishing venues and for appointment, tenure and promotion committees needing to evaluate applicants' work. Whilst some prominent rankings are citation-based, others are constructed by field experts. This article assesses whether Large Language Models (LLMs) can provide credible new business and management journal rankings.

Design/methodology/approach: Based on mean ChatGPT 4o-mini scores for business and management articles published between 2014 and 2020 and submitted to the UK Research Excellence Framework (REF) 2021, ChatGPT-based rankings were compared with expert rankings from the Australian Business Deans Council (ABDC) and the Chartered Association of Business Schools (CABS), weighted normalised citation-based rankings, mean REF citation scores, and mean REF departmental quality scores.

Findings: For the 43 journals with at least 50 articles and data from all six sources, the ChatGPT scores correlated more strongly with expert rankings (CABS: 0.438, ABCD: 0.510) than any of the citation rankings except Scimago Journal Rank (SJR) for one of the two (CABS: 0.664, ABCD: 0.360). Journal scores calculated from REF departmental quality score rankings had the highest Spearman correlations with the established rankings, however (CABS: 0.717, ABCD: 0.583). If rankings based on REF departmental quality scores are taken as optimal, then ChatGPT scores have the highest correlation with this (0.830), greater even than with the two expert rankings.

Originality/value. ChatGPT-based journal quality scores are plausible new ranking mechanism for business and management journals and may be superior to citation-based rankings in some cases, potentially providing more current, finer grained and cheaper results.

Keywords: ChatGPT, Large Language Models, research quality evaluation, journal rankings

Introduction

Although research evaluators are often encouraged to ignore the prestige of the publishing journal when evaluating research quality (Hicks et al., 2015; Wilsdon et al., 2015), business and management journal rankings continue to be influential. This is especially for new scholars choosing where to submit their papers and as a shortcut for members of appointment and promotion committees lacking the time or expertise to read publications on candidates' CVs (Anderson et al., 2022; Walters, 2022). It seems common for academics to feel forced to consult these rankings, despite misgivings about their value or accuracy (Anderson et al., 2021; Serenko & Bontis, 2024), and their potential negative systemic effects (Renwick et al., 2019). They are considered less important in some business and management specialities, perhaps partly due to these problems (Brooks et al., 2023).

The importance of rankings is also reflected by the existence of at least five for business and management based on expert judgement (Walters, 2024). For example, the Australian Business Deans Council rankings are published every three years. The 2022 version reported four quality levels (A*, A, B, or C) to 2,680 business-related journals. The process was overseen by a steering group and seven panel chairs and was a light touch process

[1] Aslib Journal of Information Management, to appear.

focused on journal name changes and journals to be added or removed, but not reranking existing journals (ABDC, 2023). Expert rankings are labour intensive, the experts might be political or otherwise not objective about journals for specialities that they do not know as well as for national journals that they are unfamiliar with (Serenko & Bontis, 2018). In addition, levels are typically relatively coarse, with no indication that a level X journal is close to the level above or below, and there may be apparent biases, such as in favour of quantitative research (Vogel et al., 2017).

Although more objective rankings could be obtained by using citation rates for journals in different ways (e.g., Moussa, 2019), and citation rates are known to correlate positively with research quality in business and management (Thelwall et al., 2023), citations have their own limitations. They may undervalue topics that are naturally less cited, nationally-focused research from smaller countries, and journals publishing research that generates longer term interest (because journal citation rate formulae typically use the most recent 2-5 years). In addition, there are many technical issues with the most used journal impact formulae (Seglen, 1998).

More fundamentally, research quality is usually conceived as comprising rigour, originality, scholarly significance, and societal significance (Langfeldt et al., 2020), with citation rates only directly reflecting scholarly influence (Aksnes et al., 2019). Expert rankings are typically informed by citation data, however, and so the two ranking types are not independent. A systematic comparison of five business/economics journal rankings with various citation-based indicators, including Scimago Journal Rank (SJR), found mostly strong correlations between them. Source weighted (i.e., citations from more important journals weighted more), size independent (i.e., without giving larger journals an advantage) indicators like SJR had the strongest correlation with expert ranks. For example, there are Pearson correlations from 0.63 to 0.72 between SJR, after a transformation, and the five expert ranks. In some cases, the citation-based indicators seemed to correlate more strongly with a journal ranking than they did with each other, although different types of correlation were reported for the two contexts (Walters, 2024). This apparent anomaly is possible since all five rankings have a national home, and some were created to address the lack of coverage of important national journals in other rankings.

Large Language Models have recently been shown to be capable at a wide range of previously specialist text processing tasks (Chang et al., 2024), so may be able to help with journal ranking. Recent research has shown that ChatGPT can be configured to assess the quality of academic journal articles after being primed with the UK Research Excellence Framework (REF) guidelines for human expert reviewers, with results that align better with human scores than citation rates and traditional machine learning (Thelwall & Yaghi, 2024; Thelwall et al., 2024). Surprisingly, it seems to produce the best quality predictions, in the sense of correlating most strongly with expert judgements, when the prompt contains just the article's title and abstract, rather than its full text (Thelwall, 2024ab). This may be due to an apparent tendency for it to give higher scores to longer documents (Kousha & Thelwall, 2024), perhaps by treating each document as a source of cumulative evidence to support the best outcome. Google Gemini seems to have a slightly weaker capability to score journal articles for research quality, although it does not seems to have a problem with longer texts (Thelwall, 2025). Most relevantly, if ChatGPT's quality scores are averaged across all articles published in a journal, then the results tend to correlate positively (median Spearman's rho = 0.62) with Finnish/Polish/Norwegian national journal ranks for the 17 largest monodisciplinary journals in the Scopus Business, Management and Accounting category (Thelwall & Kousha, 2025).

Given the potential of Large Language Models for text evaluation tasks, it is important to assess whether ChatGPT scores a useful for a wider range of business and management journals and for specialist business rankings, rather than for national rankings. This article also introduces a second and UK-based journal ranking: the average business departmental REF score of the UK-authored articles published in a journal. This exploits the UK quality scores for departments published from REF2021 (details below) together with the lists of journal articles that were used to obtain the scores.

- RQ1: Are business and management journal rankings from average ChatGPT scores more closely aligned with expert rankings than citation-based indicators from the UK perspective?
- RQ2: Are business and management journal rankings from UK authors' departmental research quality scores more closely aligned with expert rankings than citation-based indicators?

The final research question uses the perspective of selecting articles for the UK REF, for which the departmental REF scores are the best available data.

- RQ3: Are journal rankings from average ChatGPT scores more closely aligned with rankings based on UK REF departmental quality than expert rankings or citation-based indicators?

Methods

Data

From the UK business and management perspective, the most important research outputs produced are arguably those submitted to the REF. This is a systematic national science-wide assessment of the research produced by academics over the previous 6 or 7 years, with the results being used to allocate government research block funding for the next period as well as (informally) to rank departments (Pidd & Broadbent, 2015; Blackburn et al., 2024). The core of the REF is assessing the quality of submitted outputs (but see: Pinar, & Horne, 2022). In REF2021 these were an average of 2.5 articles, chapters, books, or other research projects per full time equivalent academic. In each of 34 disciplinary groupings, the outputs were scored by a collection of experts over a year. The REF results are particularly useful for the current article because the evaluators were repeatedly instructed to ignore all journal rankings and citation data (Blackburn et al., 2024), although some may have disobeyed the instruction or been unconsciously influenced by them.

The relevant UK REF 2021 Unit of Assessment (UoA) is UoA 17 Business and Management Studies. This consists mainly of journal articles published between 2014 and 2020 but also some monographs, chapters, and other outputs. These were scored by the expert reviewers with one of four quality levels 1* (nationally relevant), 2* (internationally relevant), 3* (internationally excellent) or 4* (world leading). Although the individual scores have been deleted, the number of scores at each level for each department is public. Thus, for every UK business/management school/department, the average quality of its research can be estimated by the mean of its REF 2021 output scores. These averages were used as a proxy score for the quality of each of a department's articles (see below for how they were used). As will be evident below, this substitution of departmental quality for article quality is imperfect because it is likely to reduce the strength of all article-level correlations with the data.

A list of the 15,603 journal articles submitted to UoA 17 in REF2021 was obtained from the official online source (results2021.ref.ac.uk/outputs). Duplicates (collaborative articles submitted by multiple institutions) were removed. Articles with a DOI were then queried in Scopus during November 2024 to obtain citation counts for one of the rankings and to obtain their abstracts for ChatGPT (see below). Scopus has slightly greater coverage than the Web of Science (Martín-Martín et al., 2021), so is a better choice here. The dataset was filtered to remove all articles without an abstract or with a relatively short abstract (in the bottom 10%, with under 682 characters), after removing any copyright statements. Short abstracts were more likely to be from short form submissions or editorials that would be unfair to include in journal-level calculations, and occasionally truncated abstracts presumably due to a technical error. Articles published in 2024 were also removed – these were eligible for REF2021 due to an online first publication date within the 31 December 2020 REF2021 deadline even though their formal issue publication date was afterwards. The final dataset contained 10,600 non-duplicate journal articles with DOIs and non-short abstracts.

ChatGPT scores and journal rankings

Each article was submitted to ChatGPT 4o-mini five times through its API in November 2024, and the average of the five scores used as the article's ChatGPT score. Multiple scores were used because averaging them improves the accuracy of the scores and five is sufficient for the biggest improvement (Thelwall, 2014ab). The prompt used was "Score the following article:", followed the article title, "\nAbstract\n", and then its abstract but not its full text because this approach seems to give the best results (Thelwall, 2014b). Not submitting the full text is appropriate here because the goal is to get the best ChatGPT article quality prediction rather than to get ChatGPT to evaluate the articles. The code is available online at: https://github.com/MikeThelwall/LargeLanguageModels/.

The above prompt would be insufficient for the task because it does not explain the criteria that should be used to score the article, nor the scale to be used. To address these gaps, ChatGPT was given system instructions that are a slight rewording of the official REF guidelines that the UoA 17 Business and Management evaluators had been given to guide their quality ratings for articles. In REF jargon, these are the Main Panel C guidelines (see appendix for the exact instructions: Thewall & Yaghi, 2024). The instructions essentially ask ChatGPT to judge the rigour, originality, and significance of the article and then allocate it a research quality score from 1* (nationally recognised), 2* (internationally recognised), 3* (internationally excellent) and 4* (world leading). The scores were extracted from the ChatGPT reports by text matching rules in a program written for this task (see the AI menu of: github.com/MikeThelwall/Webometric_Analyst). When the rules did not match a score in the ChatGPT report, the program displayed the report and requested a score from the user. Although ChatGPT usually gave a whole number score, fractional scores were accepted and if it gave separate scores for significance, originality, and rigour but not an overall score then the mean of these three scores was used as the overall score, including if it was a fraction.

The mean ChatGPT scores tend to be lower for older articles, perhaps because it considered older research to be less novel. To correct for this, the mean ChatGPT score was calculated for each year and then a correction factor added to each ChatGPT score, depending on the publication year, to ensure that the mean ChatGPT score was the same for all years.

For each journal, its ChatGPT score was calculated to be the mean of all corrected scores for its articles. This is labelled the *REF GPTn* score for the journal to emphasise that it is calculated exclusively from REF scores.

Journal expert rankings: ABDC and CABS

Two major business journal rankings were selected for comparison out of the five previously analysed (Walters, 2024): The Australian Business Deans Council (ABDC) Journal Quality List 2022 (abdc.edu.au/2022-abdc-journal-quality-list-released) and the Chartered Association of Business Schools (CABS) Academic Journal Guide 2021 (charteredabs.org/academic-journal-guide/academic-journal-guide-2021). The other three rankings are less than half as big as these and are based in non-English speaking nations so are less relevant.

The ABDC 2022 ranking is the 2019 ranking without changing journal ranks but adding new journals and removing old journals. This is therefore approximately contemporary with the 2014-2020 REF2021 article dataset. Journals not matched in this set were checked for in the 2019 or 2016 versions in case of name changes. The CABS 2021 rank is also approximately contemporary with the article data and was therefore used in preference to the CABS 2024 data (For concise descriptions of these sources, see Table 1 of: Walters, 2024). The two rankings could be reasonably described as citation-informed expert rankings rather than largely or completely independent of citation rate considerations. The extent of the influence of journal citation rates is unclear, however. For example, the evaluators might use one journal citation indicator as the default (e.g., using thresholds to convert to rankings) or might only consult them when a journal was unknown, a decision was marginal, or when multiple evaluators disagreed.

Journal citation rankings: SJR and REF MNLCS

The citation-based journal rankings that correlate most closely with the expert business journal rankings are those that allocate greater weighting to citations from more cited journals and calculate average rather than total citation impact (Walters, 2024). Of these the Scimago Journal Rank has the (joint) highest correlation with the ABDC and CABS rankings so was selected. The 2021 SJR iteration (www.scimagojr.com/journalrank.php?year=2021), which was published in 2021 based on earlier data, was used to be contemporary with the REF2021 article dataset. SJR 2021 measures the mean weighted citation count for documents published in the journal 2018-2020 from 2021 documents. Citations from more cited journals are given higher weightings (Guerrero-Bote & Moya-Anegón, 2012).

The second citation-based journal ranking used the Scopus citation counts obtained as above, which give least four complete years of citations for almost all articles in the dataset (except those first published late November-December 2020). This gives mature citation count data that is suitable for citation analysis (Wang, 2013).

The Scopus citation counts were not used directly for two reasons: older articles tend to be more citated and citation datasets are typically highly skewed (Baum, 2012). To avoid skewing, all citation counts were transformed with log(1+x) and to avoid year biases, the transformed values were divided by the average of the transformed citation counts from all articles in the dataset from the same year. The result is a skew-corrected ratio expressing how far above or below average an article is cited, with 1 being the mean for each year. These Normalised Log-transformed Citation Scores (NLCS) (Thelwall, 2017) are fair to compare between years. Although log-transformed indicators are not widely used, they are statistically better because, they avoid taking arithmetic means of skewed datasets. Percentile-based journal rankings could also have been used (e.g., Perianes-Rodríguez et al., 2024).

For each journal, the Mean NLCS (MNLCS) for its REF articles was calculated as its contemporary average year normalized citation rate. This is labelled the REF MNLCS to emphasise that it is only based on REF articles (unlike SJR, which uses all Scopus-indexed articles).

Journal REF quality rankings: Dept REF mean

It would be useful to directly score business and management journals from a REF quality perspective by calculating the average REF score for each article in the journal. Unfortunately, article REF scores are not published and the only information available is the proportion of REF scores at each level for each department (results2021.ref.ac.uk/profiles/units-of-assessment/17). The departmental profiles were used to calculate an average REF score for the department, treating labelled submissions from the same university as separate departments. Each article was then assigned the average REF score of its submitting department as a proxy for its original quality score. This approximation is likely to reduce the size of any underlying correlation between this average and other indicators but is at least an indirect indicator of the average quality of the UK articles, albeit biased by the REF selection decisions since academics should submit only their best work to the REF.

For each journal, the mean of the departmental REF means of the REF2021 articles published in it was calculated. This is labelled the Dept REF mean score. It most directly reflects the average quality of the UK departments submitting to a journal but is also an indirect indicator of the average quality of the UK articles submitted to it.

Analysis

The rate of agreement between the ranking schemes was assessed using Spearman correlations since some of the data consists of ranks. Although it would have been possible to transform the data to fit the Pearson correlation assumptions for statistical tests (e.g., Walters, 2024), Spearman correlations were reported for simplicity and comparability.

To make correlations fully comparable, the dataset was reduced to only articles with scores on all six indicators. This has the indirect effect of tending to remove less mainstream or smaller business journals as well as those that have little relevance to the UK and Australia.

The REF-based journal indicators will tend to be more reliable for journals with more articles due to greater averaging. Adding a threshold, such as only analysing journals with at least 10 REF articles, would solve this problem but unfortunately there isn’t a clear reason to choose any threshold value. Thus, seven journal minimum size thresholds were used (1, 5, 10, 20, 30, 40, 50) to show the extent to which the results depend on the journal size. In theory, the correlations should increase as the threshold increases due to the greater averaging, but this does not necessarily occur because increasing the threshold changes the nature of the sample, probably increasing the proportion of larger journals and those with a greater focus on the UK. Of course, the averaging issue only applies to the three REF indicators. Changes in correlations between the other three indicators are only due to changes in the sample.

Results

The analysis focuses initially on the most complete data (Table 1), then discusses the datasets with higher journal thresholds.

Table 1. Descriptive statistics for the journal articles analysed.

Year	Mean ChatGPT score	Mean Ln(1+cite)	Articles	ChatGPT offset
2014	2.949	3.784	901	0.024
2015	2.966	3.747	1255	0.007
2016	2.955	3.646	1463	0.018
2017	2.967	3.519	1546	0.006
2018	2.981	3.361	1703	-0.008
2019	2.987	3.211	1705	-0.014
2020	2.992	3.035	1491	-0.019

Analysis of all journals

For RQ1 based on all data (Table 2), journal rankings from average ChatGPT scores are **not** more closely aligned with expert rankings than citation-based indicators. The ChatGPT-based journal rankings correlate moderately (CABS: 0.398, ABDC: 0.390) with the two expert rankings, but the correlations are much stronger for SJR (CABS: 0.715, ABDC: 0.723), although they are slightly weaker for the directly comparable REF MNLCS (CABS: 0.374, ABDC: 0.334).

For RQ2, journal rankings from UK authors' departmental research quality scores are also **not** more closely aligned with expert rankings than citation-based indicators. Although the departmental REF score-based journal rankings correlate strongly (CABS: 0.675, ABDC: 0.630) with the two expert rankings, this is not quite as strong as the SJR correlations with them.

The weak correlations between the expert rankings and the three REF-based indicators in Table 2 could be at least partly due to the extra noise introduced by the smaller journals. For example, the three REF scores for a journal with one REF article would be unreliable central tendency estimates compared to the same scores for a journal with 50 REF articles because the latter estimate would be the average of 50 scores. In contrast, the expert rankings are, at least in theory, and SJR is certainly, based on all articles recently published by a journal rather than one. In practice, experts may have less knowledge of smaller journals so their opinions would presumably be less reliable for small specialist journals.

For RQ3, the correlation with the Departmental REF mean ranking is central. For this the expert rankings and Scimago Journal Rank are much better than the ChatGPT rankings for all journals (Table 2).

Table 2. Spearman correlations between the six sources of ranks for all 629 journals with non-zero scores on all ranking schemes.

1+ articles	CABS 2021	ABDC 2022	SJR 2021	REF MNLCS	REF GPTn	Dept REF mean
CABS 2021	1.000	0.758	0.715	0.374	0.398	0.675
ABDC 2022		1.000	0.723	0.334	0.390	0.630
SJR 2021			1.000	0.592	0.351	0.593
REF MNLCS				1.000	0.101	0.281
REF GPTn					1.000	0.483

Analysis of journals containing more REF2021 articles

When only considering journals with a specified minimum number of REF2021 articles (Tables 3-6) the patterns are substantially different to those from the unrestricted dataset (Table 2).

There is a gradual change as the threshold increases, so the focus will be on the most restricted dataset to illustrate the trend most clearly.

Revisiting RQ1 from the perspective of journals with at least 50 REF2021 articles (and hence larger and/or more business focused and/or more UK-focused and/or higher quality [so more likely to be selected for REF articles, when there is a choice]) gives the most sharply different results (Table 8). For this set, journal rankings from average ChatGPT scores **are** more closely aligned with the ABDC expert rankings than citation-based indicators, although they are less closely aligned than SJR for the CABS rankings. There is a surprisingly only moderately strong correlation between CABS and ABDC rankings for this set, allowing these partly conflicting results. For RQ2, large journal rankings from UK authors' departmental research quality scores **are** more closely aligned with expert rankings than citation-based indicators. Thus, for the larger journals in the set, the departmental REF mean seems like the best available journal quality indicator with ChatGPT second, and Scimago Journal Rank third.

For RQ3, the ChatGPT rankings are marginally the best for the more restricted set with at least 20 articles (Table 4) and when the set is restricted to journals with at least 50 articles, then ChatGPT rankings are clearly the best (Table 8). The more restricted set presumably contains mainly popular general journals, so it is less likely that the department journal ranking for these would include inappropriate rankings due to smaller specialist journals that some of the higher quality departments publish in due to the specialties of some of their researchers, irrespective of the quality of the research.

Table 3. Spearman correlations between the six sources of ranks for all 293 journals with at least five articles with non-zero scores on all ranking schemes.

5+ articles	**CABS 2021**	**ABDC 2022**	**SJR 2021**	**REF MNLCS**	**REF GPTn**	**Dept REF mean**
CABS 2021	1.000	0.675	0.708	0.389	0.533	0.771
ABDC 2022		1.000	0.685	0.287	0.564	0.696
SJR 2021			1.000	0.647	0.543	0.686
REF MNLCS				1.000	0.171	0.258
REF GPTn					1.000	0.657

Table 4. Spearman correlations between the six sources of ranks for all 197 journals with at least 10 articles with non-zero scores on all ranking schemes.

10+ articles	**CABS 2021**	**ABDC 2022**	**SJR 2021**	**REF MNLCS**	**REF GPTn**	**Dept REF mean**
CABS 2021	1.000	0.674	0.704	0.348	0.485	0.749
ABDC 2022		1.000	0.662	0.286	0.565	0.683
SJR 2021			1.000	0.631	0.553	0.698
REF MNLCS				1.000	0.196	0.225
REF GPTn					1.000	0.675

Table 5. Spearman correlations between the six sources of ranks for all 122 journals with at least 20 articles with non-zero scores on all ranking schemes.

20+ articles	**CABS 2021**	**ABDC 2022**	**SJR 2021**	**REF MNLCS**	**REF GPTn**	**Dept REF mean**
CABS 2021	1.000	0.650	0.679	0.312	0.508	0.758
ABDC 2022			0.590	0.285	0.525	0.634
SJR 2021			1.000	0.703	0.614	0.693
REF MNLCS				1.000	0.243	0.243
REF GPTn					1.000	0.761

Table 6. Spearman correlations between the six sources of ranks for all 86 journals with at least 30 articles with non-zero scores on all ranking schemes.

30 articles	CABS 2021	ABDC 2022	SJR 2021	REF MNLCS	REF GPNn	Dept REF mean
CABS 2021	1.000	0.686	0.663	0.348	0.478	0.747
ABDC 2022		1.000	0.586	0.336	0.523	0.654
SJR 2021			1.000	0.747	0.573	0.652
REF MNLCS				1.000	0.257	0.297
REF GPNn					1.000	0.755

Table 7. Spearman correlations between the six sources of ranks for all 59 journals with at least 40 articles with non-zero scores on all ranking schemes.

40 articles	CABS 2021	ABDC 2022	SJR 2021	REF MNLCS	REF GPNn	Dept REF mean
CABS 2021	1.000	0.555	0.625	0.353	0.403	0.708
ABDC 2022		1.000	0.514	0.338	0.470	0.568
SJR 2021			1.000	0.743	0.570	0.615
REF MNLCS				1.000	0.233	0.270
REF GPNn					1.000	0.784

Table 8. Spearman correlations between the six sources of ranks for all 43 journals with at least 50 articles with non-zero scores on all ranking schemes.

50 articles	CABS 2021	ABDC 2022	SJR 2021	REF MNLCS	REF GPNn	Dept REF mean
CABS 2021	1.000	0.526	0.664	0.404	0.438	0.717
ABDC 2022		1.000	0.360	0.245	0.510	0.583
SJR 2021			1.000	0.772	0.511	0.591
REF MNLCS				1.000	0.285	0.344
REF GPNn					1.000	0.830

The relationship between journal scores from ChatGPT and departmental REF scores for the largest journals is close to linear without major outliers (Figure 1). The minor outliers, such as Regional Studies, which has a relatively low REF score for its departmental REF mean score, could be due to departmental peculiarities affecting the latter score (e.g., a department with a particularly weak regional studies research group) or biases in the ChatGPT scores.

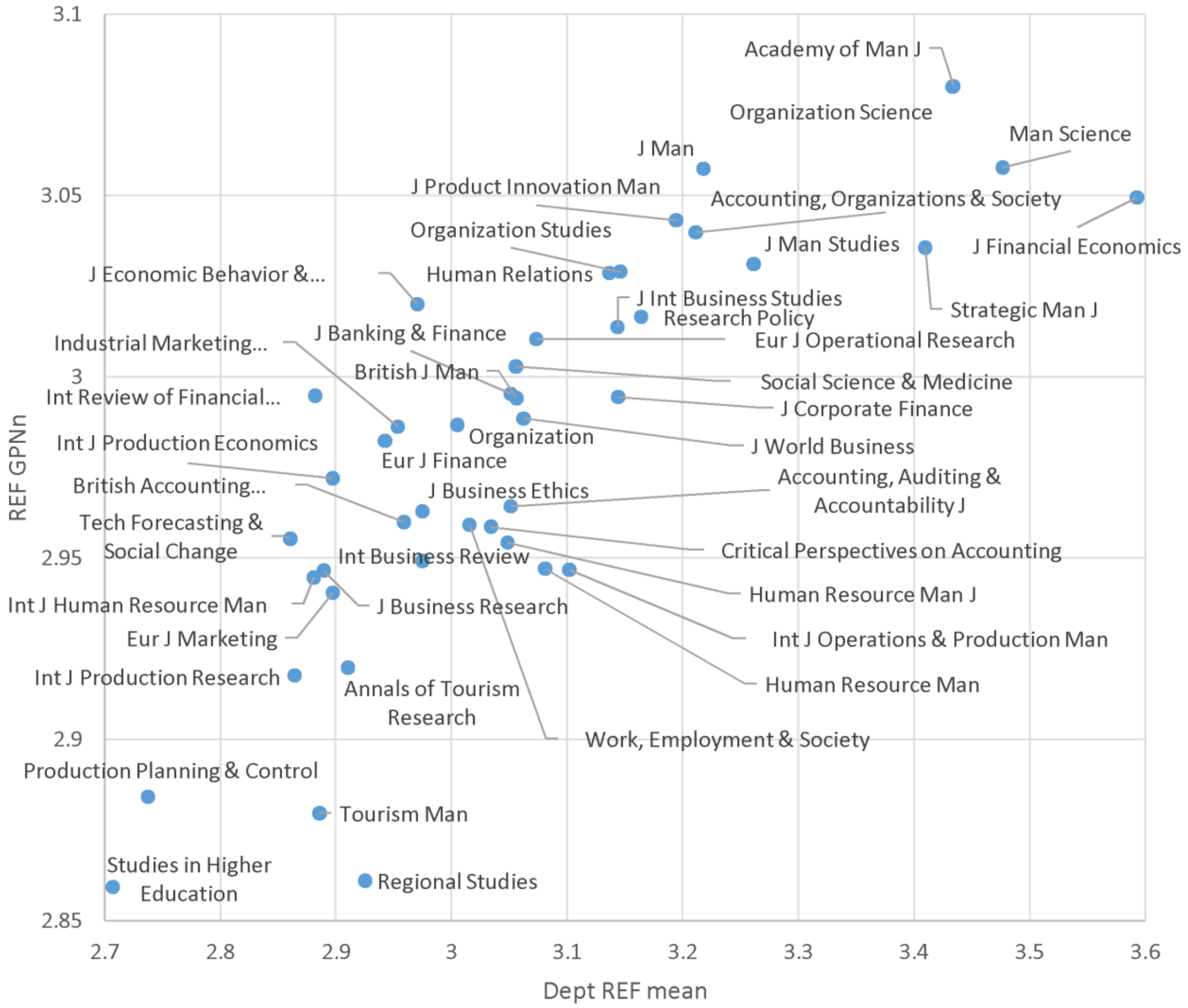

Figure 1. Average corrected ChatGPT 4o-mini score per journal against mean departmental mean REF score by journal for submitted REF2021 articles. Qualification: Journals with at least 50 REF2021 UoA 17 Business and Management articles.

Discussion

From an international perspective, the results are limited by the UK focus for the REF articles. From all perspectives, they are limited by the focus on the 43 journals most submitted to REF2021 by UK business and management scholars. It is not clear whether similar results would apply to substantially smaller journals or those with a non-UK focus. Moreover, journal importance and scope vary over time and the non-REF rankings do not exactly match the dates of the REF rankings, giving them a slight disadvantage for correlations. Recall also that the expert rankings used here are influenced to an unknown extent by journal citation rates.

There do not seem to have been any prior studies focusing on ranking business journals with ChatGPT. The results are nevertheless consistent with one multidisciplinary study that included Business Management and Accounting found a similar correlation (rho = 0.62) between national journal rankings and average ChatGPT scores based on all articles (not just the UK or REF2021) for 17 large monodisciplinary Business, Management and Accounting journals (Thelwall & Kousha, 2025).

From a largely unrelated perspective, the Spearman correlation between departmental REF means and CABD ranks of rho=0.711 is higher than the Pearson correlation between article REF scores and CABD ranks (calculated by REF evaluators with access to raw data) of r=0.466 previously reported (Blackburn et al., 2024). This may be an aggregation effect, with average scores being more reliable (as well as finer grained) than individual article scores.

The results extend previous much larger scale comparisons of journal rankings which found that size-independent, weighted citation metrics to be the optimal alternative to expert rankings (Walters, 2024; see also: Walters, 2017), by introducing a ChatGPT alternative that has a higher correlation with one of the two largest rankings, and a departmental REF method that has a higher correlation than both expert rankings. Whilst the latter is limited to journals extensively used in the UK REF, the former could potentially be expanded to all journals with abstracts, giving similar coverage to citation-based indicators. This expansion should be examined cautiously, however, in case the pattern for larger journals does not extend to smaller ones.

Conclusion

The results suggest that ChatGPT ranking business and management journals, based on the average score given to article titles and abstracts, is a reasonably effective method to rank larger business and management journals and may be more effective than citation-based rankings. This is despite the ChatGPT scores being guesses based on limited information rather than proper full-text evaluations in any sense. Compared to citations (which are also limited by only incorporating one type of problematic scholarly impact data), the ChatGPT approach has the advantage that it can include current research and does not need to wait three years for sufficient citations to accrue for a robust indicator. Nevertheless, the results are not conclusive because they focus only on large journals and take a UK perspective. Assessments of all ABDC and CABD journals based on all recent articles in them are needed to check if the findings extend to all the main business and management journals.

This article has not considered whether the data sources could be combined to give results that would be more informative. For example, a weighted average of the ChatGPT score and the citation rate of a journal might be an improvement for some weights. Nevertheless, since they have different strengths and weaknesses, it seems preferable to employ them separately rather than as a more opaque hybrid indicator.

From the UK REF perspective, averaging ChatGPT scores for journals based on REF articles seems to give an accurate indicator of the mean quality of the UK departments submitting to a journal, suggesting that it is also a reliable indicator of average journal REF article quality. This might be exploited to support future REF selection decisions, especially for newer journals not considered in the previous REF and for journals that are believed to have recently changed.

A possible practical use of ChatGPT rankings might be to replace expert rankings in the longer term, if additional research verifies the suggested properties above and the results are considered credible by the academic community. Alternatively, they might be used as an additional quantitative indicator alongside citations to help expert decisions about the rankings. Their advantages might be for newer journals without mature citation data or for journals in low citation specialties. Of course, all journal rankings are problematic and should be used for research evaluations only when article-level assessments are impractical or undesirable.

References

ABDC (2023). 2022 Journal Quality List Review Final Report. https://abdc.edu.au/wp-content/uploads/2023/03/ABDC-2022-Journal-Quality-List-Review-Report-150323.pdf

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. Sage Open, 9(1), 2158244019829575.

Anderson, C. G., McQuaid, R. W., & Wood, A. M. (2022). The effect of journal metrics on academic resume assessment. Studies in Higher Education, 47(11), 2310-2322.

Anderson, V., Elliott, C., & Callahan, J. L. (2021). Power, powerlessness, and journal ranking lists: The marginalization of fields of practice. Academy of Management Learning & Education, 20(1), 89-107.

Baum, J. A. (2012). The skewed few: does "skew" signal quality among journals, articles, and academics? Journal of Management Inquiry, 21(3), 349-354.

Blackburn, R., Dibb, S., & Tonks, I. (2024). Business and management studies in the United Kingdom's 2021 research excellence framework: Implications for research quality assessment. British Journal of Management, 35(1), 434-448.

Brooks, C., Schopohl, L., & Walker, J. T. (2023). Comparing perceptions of the impact of journal rankings between fields. Critical Perspectives on Accounting, 90, 102381.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, *15*(3), 1-45.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. Journal of Informetrics, 6(4), 674-688.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature, 520(7548), 429-431.

Kousha, K., & Thelwall, M. (2024). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. arXiv preprint arXiv:2410.19948.

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. Minerva, 58(1), 115-137.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics, 126(1), 871-906.

Moussa, S. (2019). Is Microsoft Academic a viable citation source for ranking marketing journals? Aslib Journal of Information Management, 71(5), 569-582.

Perianes-Rodríguez, A., Mira, B. S., Martínez-Ávila, D., & Grácio, M. C. C. (2024). Real influence: A novel approach to characterize the visibility of journals and publications. Quantitative Science Studies, 5(3), 778-804.

Pidd, M., & Broadbent, J. (2015). Business and management studies in the 2014 Research Excellence Framework. British Journal of Management, 26(4), 569-581.

Pinar, M., & Horne, T. J. (2022). Assessing research excellence: evaluating the research excellence framework. Research Evaluation, 31(2), 173-187.

Renwick, D. W., Breslin, D., & Price, I. (2019). Nurturing novelty: Toulmin's greenhouse, journal rankings and knowledge evolution. European Management Review, 16(1), 167-178.

Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. Acta Orthopaedica Scandinavica, 69(3), 224-229.

Serenko, A., & Bontis, N. (2018). A critical evaluation of expert survey‐based journal rankings: The role of personal research interests. Journal of the Association for Information Science and Technology, 69(5), 749-752.

Serenko, A., & Bontis, N. (2024). Dancing with the devil: the use and perceptions of academic journal ranking lists in the management field. Journal of Documentation, 80(4), 773-792. https://doi.org/10.1108/JD-10-2023-0217

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. Journal of informetrics, 11(1), 128-151.

Thelwall, M. (2024a). Can ChatGPT evaluate research quality? Journal of Data and Information Science, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2024b). Evaluating research quality with large language models: an analysis of ChatGPT's effectiveness with different settings and inputs. https://arxiv.org/abs/2408.06752

Thelwall, M. (2025). Is Google Gemini better than ChatGPT at evaluating research quality? Journal of Data and Information Science, 10(1), 1–5. https://doi.org/10.2478/jdis-2025-0014

Thelwall, M., Jiang, X., & Bath, P. A. (2024). Evaluating the quality of published medical research with ChatGPT. arXiv preprint arXiv:2411.01952.

Thelwall, M., & Kousha, K. (2025). Journal Quality Factors from ChatGPT: More meaningful than Impact Factors? *Journal of Data and Information Science*. https://doi.org/10.2478/jdis-2025-0016

Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P. & Levitt, J. (2023). In which fields are citations indicators of research quality? Journal of the Association for Information Science and Technology, 74(8), 941-953. https://doi.org/10.1002/asi.24767

Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. https://arxiv.org/abs/2409.16695

Vogel, R., Hattke, F., & Petersen, J. (2017). Journal rankings in management and business studies: What rules do we play by? Research Policy, 46(10), 1707-1722.

Walters, W. H. (2017). Do subjective journal ratings represent whole journals or typical articles? Unweighted or weighted citation impact? Journal of Informetrics, 11(3), 730-744.

Walters, W. H. (2022). Evaluating journals in business and related fields: A guide for faculty. Business Information Review, 39(3), 90-97.

Walters, W. H. (2024). Relationships between expert ratings of business/economics journals and key citation metrics: The impact of size-independence, citing-journal weighting, and subject-area normalization. The Journal of Academic Librarianship, 50(4), 102882.

Wang, J. (2013). Citation time window choice for research impact evaluation. Scientometrics, 94(3), 851-872.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., & Johnson, B. (2015). The metric tide. Report of the independent review of the role of metrics in research assessment and management. https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/