WPOM

Working Papers on Operations Management.
Vol. 14, Nº2 (137-166)
ISSN: 1989-9068

https://doi.org/10.4995/wpom.19631

A noise audit of the peer review of a scientific article: a WPOM journal case study

Tomas Bonavia[a] and Juan A. Marin-Garcia[b]

[a] Department of Social Psychology, University of Valencia, Valencia, Spain. tomas.bonavia@uv.es, [b] ROGLE-Departamento de Organización de Empresas, Universitat Politecnica de Valencia, Valencia, Spain. jamarin@omp.upv.es

Recibido: 2023-04-28 Aceptado: 2023-07-10

Abstract

A Spanish version of the article is provided (see section before references).

This study aims to be one of the first to analyse the noise level in the peer review process of scientific articles. Noise is defined as the undesired variability in the judgements made by professionals on the same topic or subject. We refer to evaluative judgements in which experts are expected to agree. This is what happens when we try to judge the quality of a scientific work. To measure noise, the only information needed is to have several judgements made by different people on the same case to analyse their dispersion (what Kahneman et al. call a noise audit). This was the procedure followed in this research. We asked a set of reviewers from the journal WPOM (Working Papers on Operations Management) to review the same manuscript which had been previously accepted for publication in this journal, although the reviewers were unaware of that fact. The results indicated that if two reviewers were used, the probability of this manuscript not being published would be close to 8%, while the probability of it having an uncertain future would be 40% (one favorable opinion and one unfavorable opinion or both suggesting substantial changes). In the case of employing only one reviewer, in 25% of the cases, the audited work would have encountered significant challenges for publication. The great advantage of measuring noise is, once measured, it is usually possible to reduce it. This article concludes by outlining some of the measures which can be put in place by scientific journals to improve their peer review processes.

Keywords: peer review; evaluation of scientific journals; research evaluation; decision making; decision noise; making judgements

Introduction

Since the 2021 publication of the book by Kahneman, Siboni, and Sunstein, "Noise: A Failure of Human Judgement", and its prequel (Kahneman et al., 2016), a new topic has emerged having significant impact. By early 2023 more than 65 scientific articles indexed in SCOPUS cited the work of Kahneman et al. (2016), which placed this article in the first percentile of the most-cited articles (by year and by area). However, as far as scientific research and, specifically, peer review processes are concerned, although much has been published thus far (Bornmann, 2011; Weller, 2001), it has not been published in the perspective of analysing the noise level. This article intends to be an exception to the above situation. In fact,

we believe it is the first time a scientific journal has embarked on an audit to assess the noise level in its peer review processes.

Noise in this context is defined as undesirable variability in judgments about the same problem (Kahneman et al., 2021, p. 19, emphasis added). Variability in judgments is not always undesirable such as when it comes to matters of opinion, preference, or taste. But one should not confuse personal taste with professional judgement. For example, it is one thing for two people, even two specialists, to differ on which painting they like best, but it is quite another thing to establish its price when appraising it. The word "judgement" is mostly used when it is considered people must agree (Kahneman et al., 2021). With matters of opinion or taste differences are acceptable, even desirable, which differs from matters of judgement. In matters of judgement what we expect is that experts tend to agree and differences are minimised. If you have an illness, for example, you will not be happy to receive two different diagnoses, not least because you will rightly think at least one of the two must necessarily be wrong. If you are being assessed on your performance in your company, you will not understand how two human resources professionals can make different judgements about your actual level of performance. There are countless examples where excessive variability in judgements would not be considered acceptable by most (e.g., in determining the disability degree of a person, the criminal sentence a defendant should receive, or the university entrance exam grades of your children, etc.).

It is true in scientific publishing it is assumed the same manuscript can be assessed differently by different reviewers (Benda and Engels, 2011; Theoharakis et al., 2007). Many journals state in their instructions to authors that they will submit manuscripts to two anonymous reviewers and, if there is no agreement between them, they will submit it to a third, with the journal editor always having final say. This is common practice in all scientific disciplines, not only in the social sciences. This means nothing more than assuming in science there is no single criterion for validating the quality with which a paper has been produced, or, even if there is some consensus on criteria, they are not necessarily applied equally by the researchers themselves (Bedeian, 2004). As Kahneman et al. (2021) stated, accepting a certain level of variability is tolerable when the decision-makers are human beings. But we must not forget we are talking about judging the quality of scientific work, not just a matter of opinion. It is a decision akin to one taken by a judge in their sentencing (it is not expected that, faced with the same crime, two judges can issue different sentences). Thus, we can ask, is it fair the same manuscript submitted to a journal can be assessed so differently, e.g., ranging from outright rejection to intermediate assessments suggesting modifications of notable degrees of variability to acceptance with minimal changes?

In summary, we can assume evaluative judgments involve an expectation of limited disagreement. As Kahneman et al. (2021) argued, evaluative judgments depend, in part, on the values and preferences of those who make them, but they are not mere matters of taste or opinion. Great variability in judgements about the same case should always be undesirable. Whenever the person making a judgement is randomly selected, or nearly so, from a pool of equally qualified people (as is the case with the peer review process in science), noise is a problem. The above implies a logical assumption of a certain degree of randomness in the choice of articles which are eventually published. Publication can even become a lottery (depending on which reviewer you end up with, the article may or may not see the light of day). We start from the erroneous assumption that another expert would make a similar judgement on the same stimulus (e.g., a scientific manuscript), but this is often not the case (Ernst et al., 1993). In different studies conducted in different settings (court rulings, insurance agents, etc.), Kahneman et al. (2021) found noise, on average,

is five times greater than one might initially think. In the specific case of peer review, agreement between judges, when corrected for probability, ranged between 0.20 and 0.40, which corresponded to a low level of agreement (Bornmann, 2011).

We propose a first approach to this issue from a different perspective. Peer review is an evaluative judgement which, by its nature, is unverifiable. However, all we need to measure noise are multiple judgments about the same problem, we do not need to know an actual value (Kahneman et al., 2021, p. 27). The only information required is to have several judgements made by different people about the same case and to analyse their dispersion in order to see what values they reach. Noise audit is the name Kahneman et al. used to refer to this procedure. What do we need to start? We need different referees to make judgements on the same case. For this article we needed different referees to review the same manuscript submitted to a scientific journal for publication, and then to analyse the degree of variability in their judgements. Moreover, once the noise was measured, it was generally possible to reduce it. This was what Kahneman et al. (2021) called decision hygiene, a concept that referred to the set of strategies and techniques which can be employed to reduce noise.

Method

In early 2022 we chose for this study the last accepted but unpublished manuscript in the journal Working Papers on Operations Management (WPOM). The manuscript had completed the normal review process for the journal, i.e., two reviewers assessed the paper in a double-blind format and the editor-in-chief made the editorial decision to accept it. We made the paper selection prior to approaching the editor-in-chief (author 2) about collaborating on our noise audit research.

The authors of the manuscript were asked via email for their informed consent for their work to be included in the study (see Appendix 1). We received agreement from all authors by 16 February 2022.

The editor-in-chief of the journal selected 24 reviewers from among their most active and timely reviewers. All held PhDs with years of experience, had stable positions in academia, and published articles in internationally recognised journals. The journal used a standard invitation email without adding any additional information about our audit and they were given the standard deadline (i.e., one week to accept or decline the assignment and 4 weeks to complete the review). Their participation, as per journal procedure, was confidential and all data were anonymised before being processed during the research. Therefore, the review process was blind during the noise audit.

In total 18 participants decided to accept the review request. All of them used the official journal review template for systematic literature review protocols (the type of article used for this noise audit test). The items of this review template were based on the 2009 PRISMA statement, among others.

The authors of the manuscript selected for this audit received a copy of all the comments from the 18 extra reviewers. These authors valued these comments and took advantage of them to improve their own research, which was eventually fully published (Álvarez and Maheut, 2022). Therefore, we can affirm the extra review work carried out benefited to produce improved published science.

Inter-judge reliability is typically measured with the Krippendorf alpha coefficient and similar measures (Krippendorff, 2011; LeBreton and Senter, 2008). The higher the alpha coefficient, the lower the noise (an alpha value of 1 reflects full agreement). In this audit, having only one object to be evaluated by 18

judges, it was not possible to use these measures of inter-judge reliability (for a review of these see, for example, Benda and Engels, 2011; or Weller, 2001), so we have analysed the concordance (degree of agreement) among the participants in the review process and performed a probability analysis, estimating the percentage of equal rankings against the number of possible ranking alternatives.

Results

We begin the description of the results of our study with the overall assessment by the reviewers of the manuscript we asked them to evaluate.

In viewing Table 1 below it should be borne in mind that, as were 4 response alternatives, the maximum noise level would have been obtained if each alternative were selected in 25% of the cases, i.e., maximum dispersion. The minimum noise level would have been obtained when only one of the four alternatives was selected in 100% of the cases, i.e., maximum level of concordance.

Table 1. Final recommendations of the reviewers for the submitted manuscript

	Frequency	Percentage
Refuse submission	2	11.1
Major changes required	3	16.7
Minor changes required	10	55.5
Accept submission	3	16.7
Total	18	100.0

In the review of scientific articles, given the review process follows a prior assessment by a scientific editor of the journal, the most likely response option is the reviewers request some level of changes from the authors. These changes may be minor or may involve major modifications to the manuscript (the terms commonly used are minor revision and major revision). In WPOM the same distinction was followed (minor and major changes), in addition to rejection or outright acceptance of the manuscript. The usual decision process in WPOM was as follows: (a) if all reviewer opinions are favourable (accept as-is or minor changes) the paper is accepted for publication and the final version will only be reviewed by the editor in charge of the submission; (b) when all opinions are negative (reject the submission) the editor-in-chief assesses the criticisms made by the reviewers and, unless the reviewer opinions are not properly justified, the article will be rejected; c) in any of the other cases, authors are offered the option of going on to a new round of review, although, if they have received major criticisms from any of the reviewers, the chances of publication were drastically reduced.

Given the manuscript used in this audit had already been accepted (as described in the Procedure section), in case of an absence or a low noise level, a clear majority of reviewers should have agreed on the same alternative aligned with the editorial decision. In other words, the submission recommendations should have been between accepting or introducing minor changes. However, as can be seen in Table 1, the option with the highest percentage (minor changes) was just over 50%. In total, 13 of the 18 reviews (72%) agreed with the editorial decision already taken. The remaining 28% opted for a clearly different assessment (suggesting major changes or recommending outright rejection). This meant on 1 out of 4 occasions this manuscript would have had many problems for publication or would not have been published direct-

ly, and all this considering it was the same manuscript which had already followed the usual peer review process of any scientific journal.

Probability analysis

Each reviewer could choose between 4 recommendations (rejection, major changes, minor changes, and acceptance). The number of combinations with repetition of n elements taken from k by k can be calculated with the formula:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Considering there were two reviewers (a common situation in the standard review process of a journal) and the order of the recommendations was not relevant (i.e., being reviewer 1 or 2 did not give more weight to the recommendation) we found, in our case, the number of combinations of 4 elements with repetition taken in pairs was 10, the other 6 options being a situation analogous to one of these 10. Table 2 shows the combination options and the different situations in which a manuscript may be left after peer review in WPOM (in other journals it may be different).

Table 2. Different possible combinations after a peer review process with 4 assessment alternatives

Reviewer 1↓ / Reviewer 2→	Refuse submission	Major changes	Minor changes	Accept submission
Refuse submission	Reject	Difficult to publish	Undecided	Undecided
Major changes	Difficult to publish	Difficult to publish	Undecided	Undecided
Minor changes	Undecided	Undecided	Publish	Publish
Accept submission	Undecided	Undecided	Publish	Publish

Taking as a reference the use of two people as reviewers (the most usual situation for a journal), if the recommendations were evenly distributed (same probability for each of the options, equal to 0.25), each of the boxes in Table 2 would have a 6.25% probability (0.25*0.25). In that case, considering the usual decision process in WPOM (as described in the previous section), the probability of a direct rejection would be 6.25%, a new round of review which would be difficult to publish would be equivalent to 18.75%, 50% of the manuscripts would have an uncertain future, and the remaining 25% would be published with certainty (either directly or after simple and quick changes to the original manuscript). In the no-noise case, i.e., if the two reviewers always agreed, only the cells on the diagonal would be possible (each with a 25% probability), so that the probability of rejection would be 25%, the probability of a new round of difficult-to-publish review would also be 25%, and 50% of the manuscripts would be published (no article would have an uncertain future).

However, the articles entering the review process passed through the previous filter of the editor-in-chief and it did not seem reasonable the four options for recommendation were comparable. And if, as in our case, the manuscript was already accepted for publication, even less so. Therefore, it would be expected

that the reviewers who participated in our study, in conditions of absence or of low noise, would have tended to coincide in determining their assessment was to publish it.

Extending the analysis of Table 2 to the situation of this audit with 18 reviewers, 153 different pairs of reviewers could have resulted. Furthermore, instead of assuming a uniform distribution of reviewer assessments, we used the actual frequencies of decisions made. In this case, the probability of a direct rejection was 1.2% (probability that, in the total of 153 pairs, two people agree to reject the manuscript), the probability that a new round of review was proposed which is unlikely to result in publication was 6.5%, 40.1% of the manuscripts would have an uncertain future, and the remaining 52.1% would be published with a high degree of certainty. In other words, with a 50% probability, the possible random combinations of people who have participated in this noise audit would have aligned with the editorial decision to accept the manuscript, which implied the remaining 50% could have been considered noise.

Detailed analysis of the peer review

The above results referred to the final decision which could have been made about a peer-reviewed article. On the other hand, the WPOM journal used a template with different items to evaluate each of the manuscripts adapted according to the type of work it received (for this audit it was a systematic review of the literature). Table 3 shows the breakdown of the quantitative scores assigned to each of the items by the people who reviewed that manuscript.

Table 3. Frequency and percentage (in brackets) of the variables studied for each of the response alternatives

Number	Item	not at all	just marginally and needs to be improved	yes (it is sufficient)	Not applicable, or no answer
1 SLRjam01	Does it contain all the parts of a literature review protocol?	0 (0)	5 (27.8)	13 (72.2)	---
2 SLRjam02	Is it understandable by someone who is not an expert?	1 (5.6)	4 (22.2)	13 (72.2)	---
3 SLRjam03	Are all the variables properly defined?	1 (5.6)	4 (22.2)	10 (55.6)	3 (16.6)
4 PRISMA03	Does it describe the rationale for the review in the context of what is already known?	1 (5.6)	5 (27.8)	12 (66.6)	---
5 PRISMA04	Does it provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design?	1 (5.6)	3 (16.6)	14 (77.8)	---
6 SLRTemp2d	If extending previous research on the topic, does it explain why a new study is needed?	2 (11.1)	7 (38.9)	9 (50)	---
7 SLRTemp3a	Specify and justify basic strategy: manual search, automated search, or mixed	0 (0)	1 (5.6)	17 (94.4)	---
8 PRISMA06 SLRTemp4a-b	Identify the inclusion criteria for primary studies, identify the exclusion criteria	0 (0)	1 (5.6)	17 (94.4)	---
9 PRISMA07 SLRTemp3c	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched	0 (0)	2 (11.1)	16 (88.9)	---

10 PRISMA08 SLRTemp3b	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	0 (0)	0 (0)	18 (100)	---
11 SLRTemp3d	For manual searches, identify the journals and conferences searched	0 (0)	0 (0)	8 (44.4)	10 (55.6)
12 SLRTemp3e	Specify the time period covered by the review and any reasons for your choice	1 (5.6)	8 (44.4)	9 (50)	---
13 SLRTemp3f	Identify any ancillary search procedures, e.g., asking leading researchers or research groups, or accessing their web sites; or checking reference lists of primary studies	4 (22.2)	0 (0)	5 (27.8)	9 (50)
14 SLRTemp3g	Specify how the search process was evaluated, (e.g., against a known subset of papers, or against the results from a previous systematic review)	1 (5.6)	3 (16.6)	10 (55.6)	--- 4 (22.2)

For the items analysed in Table 3, the highest noise level would have been obtained if 33% of the reviewer responses had been equally distributed among the three possible alternatives in each variable, apart from items 3, 11, and 13, where the maximum dispersion percentage would have been 25% due to the existence of 4 response alternatives. The results show the maximum level of agreement was reached in variable 10, and it was also quite high in variables 7, 8 , and 9. It was relatively low in variable 1, increased some in items 2, 3, 4, and 5, while it was already higher for variables 6, 12, 13, and 14. One might be tempted to think the reviewer who indicated the option "not at all" in variable 2 was the same reviewer who chose the same alternative in variables 3 and 4, but this was not the case, they were all different reviewers. To give another example, if we compare what happened with variables 7, 8, and 9, none of the reviewers who stood out from the overwhelming majority were the same. In summary, there was not always an unequivocal pattern in the choice made by the reviewers of this manuscript who chose one option in one item frequently while changing it in another such that the answers given were combined in different ways.

For the first item, when asked whether the manuscript contained all parts of the protocol for the literature review, a large majority of reviewers chose to acknowledge this criterion was sufficiently met and no one considered the opposite to be true. However, nearly one third felt the criterion was marginally met and needed to be improved. As for the comments made by the reviewers on this point, we found a range (see Table 4) from those strongly in favour (e.g., "The article is well structured and follows the usual structure of this type of article" or "The systematic literature review protocol is rigorous and follows the steps established in reference works"), to those showing the opposite (e.g., "The structure of the article is not sufficiently well structured"), to those showing the opposite but less severe (e.g., "The structure of the article does not follow the usual structure of this type of article" or "The structure of the article is not perfectly in line with a literature review protocol" or "The literature review contains only the basic and most typical descriptive analyses"). Of course, there were also comments of a more intermediate nature (e.g., "Although the required sections are there, it is suggested to also mention how you have carried out the subsequent coding process and the inclusion of a tentative work plan").

In relation to item 2 on whether the manuscript was understandable to a non-expert, the results were quite like the previous variable, except in this case one reviewer considered this criterion as not met at all. The same happened with the comments, being quite positive in some cases (e.g., "It is perfectly understanda-

ble. It is easy to read and provides an adequate knowledge of the technique and its objectives" or "The paper is written correctly"), quite negative in others (e.g., "Although the content of the article is very basic, it has serious problems of writing, which does not make it easy to read"), or showing various nuances (e.g., "Although the text can be understood by someone who is not an expert, it presents spelling, formatting and punctuation errors, which should be corrected").

We could continue describing the remaining variables one by one, but we have summarised them in Table 4 which includes comments made by the participants who reviewed the manuscript. In preparing Table 4, special care was taken to include comments from all reviewers whenever possible, however, not all reviewers included comments on each variable. Furthermore, a special effort was made to include comments from all reviewers, not just those who had a strongly favourable or unfavourable opinion of the manuscript. Additionally positive and negative comments from different reviewers were combined and seemed to contrast with each other (e.g., one review highlighted one item as positive and another item as negative, while another review highlighted just the opposite).

Table 4 shows there were suggestions which were not framed in the same way even though they were being asked about the same point and comments which were manifestly at odds with each other. It was not simply that reviewers commented on different aspects of the manuscript, as has been argued to justify low inter-judge reliability (Fiske and Fogg, 1990). Thus, in what way should have the authors of the revised manuscript improved their work, how could they have integrated suggestions that were fundamentally putting forward contradictory ideas? It seemed clear, depending on chance (bearing in mind it was routine to use two anonymous experts to review a manuscript who were chosen from a list of potential reviewers), the final published article could turn out to be quite different, or not even published at all. As Bornmann (2011) pointed out, few studies have investigated the real reasons behind reviewer disagreement (e.g., by conducting comparative analyses of the content of review sheets) and our work made a new contribution in this direction.

Table 4. Reviewer comments on the manuscript

Number	Item	Examples of positive comments	Examples of negative comments
1 SLRjam01	Does it contain all the parts of a literature review protocol?	The article is well structured and follows the usual structure for this type of article (review 1).	The structure of the article is not perfectly in line with a literature review protocol (review 5).
2 SLRjam02	Is it understandable by someone who is not an expert?	The paper is correctly written (review 5).	Although the content of the article is very basic, it has serious editorial problems, which does not make it easy to read (review 11).
3 SLRjam03	Are all the "variables" properly defined?	The variables used for the analysis are appropriate (review 11).	The article does not define the variables (review 8).
4 PRISMA03	Does it describe the rationale for the review in the context of what is already known?	The authors add a specific section to analyse the current state of literature reviews on the subject (review 5).	The authors should do more to explain why this literature review is necessary and how these results could be useful for future research (review 4).

5 PRISMA04	Does it provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design?	The authors pose some research questions and the context for the literature review is made explicit (review 4).	I understand that a detailed explanation of all these elements is not given (review 2).
6 SLRTemp2d	If extending previous research on the topic, does it explain why a new study is needed?	The authors show three articles based on literature reviews and explain them (review 8).	It is not clear what this study will add to existing studies (review 10).
7 SLRTemp3a	Specify and justify basic strategy: manual search, automated search, or mixed.	I understand that this point is explained in the article (review 2).	Clarification is requested on some of the steps in the attached document (review 9).
8 PRISMA06 SLRTemp4a-b	Identify the inclusion criteria for primary studies, identify the exclusion criteria.	The inclusion and exclusion criteria are explicit and seem reasonable (review 4).	Only listed, more detail should have been given (review 2).
9 PRISMA07 SLRTemp3c	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	The article describes the information sources used for the information search (review 11).	See improvement comments in attachment (review 3) – NB numerous comments were included in the attachment.
10 PRISMA08 SLRTemp3b	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Yes, articles that are repeated in the databases used are removed (review 5).	See improvement comments in attachment (review 3) – NB numerous comments were included in the attachment.
11 SLRTemp3d	For manual searches, identify the journals and conferences to be searched.	Yes, specified (review 2).	Not applicable (review 11).
12 SLRTemp3e	Specify the time period to be covered by the review and any reasons for your choice.	The time period was 2012 to 2020 (review 4).	The rationale for the time horizon considered for the literature review is not specified (review 14).
13 SLRTemp3f	Identify any ancillary search procedures, e.g., asking leading researchers or research groups, or accessing their web sites; or checking reference lists of primary studies.	It may be interesting to analyse the bibliography of the 40 selected articles in order to confirm the analysis strategy and incorporate important articles (review 7).	No snowball search (review 16).
14 SLRTemp3g	Specify how the search process is to be evaluated, (e.g., against a known subset of papers, or against the results from a previous systematic review).	We compare the results obtained with previous reviews (review 9).	No comparison or discussion with the results of similar systematic reviews (review 14).

Discussion

In this study, we chose to send the reviewers a manuscript already accepted by the journal editor after a prior peer review process in which modifications were made to the initial version of the manuscript. This fact should have greatly reduced the noise in the evaluation of the manuscript, as it was nominally an improved article already accepted for publication as it had been. And this was partly the case. Nevertheless, the results indicated noise existed and was not negligible. In principle, the reviewers who distanced themselves significantly from the acceptance of the article caused noise (in this case, rejections and major

revisions account for more than 25%). In other words, one out of four reviewers would not have suggested publishing this manuscript which had already gone through a standard review process and had been approved for publication. Given the level of agreement among reviewers is generally not much higher than would be expected to occur based on chance alone (Benda and Engels, 2011, p. 169), the peer review process followed by WPOM reduced this percentage (assuming that the "correct" response was to accept the article for publication).

We considered the same scientific work, a manuscript, could not be bad, good, and fair all at the same time. Similarly, it was unfair to treat authors differently, who may have received a favourable or unfavourable assessment depending on the reviewer, when both reviewers judged the same exact manuscript. A system in which professional judgements are clearly inconsistent loses credibility (Kahneman et al., 2021, p. 167). What is important in review processes is to evaluate the research work being presented. However, our study clearly showed the evaluators played a key role. This was something we already strongly suspected, but what became relevant was there was more evidence their role was decisive in judging the quality of a contribution, and probably its outcome, to the extent that in some cases their role transcended the contribution itself, leaving the latter in second place. Whether or not an article is accepted for publication by a scientific journal can sometimes run the risk of becoming a lottery, a matter of luck (Bedeian, 2004). As Kahneman et al. (2021, p. 27) state in any system in which judges are assumed to be interchangeable and assigned in a quasi-random fashion, major disagreements about the same case undermine expectations of fairness and consistency. This is a problem that fully affects scientific objectivity.

Let us begin, then, to diagnose some problems in peer review processes and propose possible solutions. To begin with, the response scale itself is a source of noise. People may differ in their judgements, not because they fundamentally disagree, but because they assign different meanings to the values on the scale. The response scale used in Table 3 did not seem likely to pose many problems in this respect as it was a priori clear (or at least the consequences of choosing one or the other option were not decisive). The same was not true, in our opinion, of the scale used to make an overall judgement about the quality of the manuscript (see Table 1). In this case, the difference between suggesting "minor revision" or "major revision" to a manuscript may not have been obvious. In fact, common experience shows what one reviewer considers minor changes are considered major changes by other reviewers, and vice versa. A clear guideline from journals as to what each of them considers to be minor or major changes could be very useful in the review process because it would reduce ambiguity, as ambiguous scales are known to be noisy (Kahneman et al., 2021).

In the specific case of WPOM, reviewers received these additional clarifications, although they may not have been sufficient: a) for minor changes; "Revisions will not be subject to a new round of peer review", and b) for major changes; "Revisions will be subject to a new round of peer review". In other cases, journals could choose to specify what they mean by minor changes, for example, by reserving such changes for editorial, formatting, structural, etc. issues. And specify that major changes are necessary when it is necessary to modify theoretical, methodological, focus, or aspects which have to do with the discussion or conclusions of the study. Each journal could determine what it considers appropriate in the different response alternatives it offers for each criterion. As reviewers often evaluate papers from different journals, they would know best what to look for in each journal.

On the other hand, variations in the noise level may be due to different reasons, not only because the reviewers have a different opinion on the degree of compliance with the review criteria established by the journals (see Table 3 for these criteria in the specific case of WPOM). It may also be because the wording of the criteria items is not always clear and can be interpreted in different ways. Reviewers may doubt what they are being asked, making it difficult to understand the variables and consequently increasing the noise level. Or worse, they may not doubt but this does not prevent different reviewers from interpreting the same criterion differently because it is also fundamentally ambiguous (ambiguity can be related both to the values of the response scale to an item, as discussed above, and to the wording of the item itself). If this is the case for some of the variables, the solution to reduce the noise level would be to clarify the wording of the questions asked of reviewers. There are standard reporting guidelines (PRISMA, STROBE, MOOSE, SQUIRE, et al.) that some journals advise their reviewers to use (although in a study by Hirst and Altman, 2012, only 35% of journals provided online instructions of some kind to their reviewers). In any case, if these guidelines were better specified and their ambiguity reduced (it would be possible to test how they are interpreted by the most common reviewers of a journal, which would mean an even more robust system), the peer review process would gain in objectivity.

Based on the results of this study, the lack of clarity in the formulation of items 3, 11, 13, and 14 could have been the cause of their higher noise level. In view of this situation, WPOM could consider reformulating the statements to make them clearer (this journal used PRISMA, although it still did not seem sufficient), adding some very specific additional information to clarify the meaning of what is being asked. It is well understood one of the biggest problems causing noise is the lack of adequate and, above all, agreed upon terminology (Kahneman et al., 2021). The solution is to ensure a common frame of reference. Guidelines can therefore be a powerful mechanism for reducing both bias and noise, because they directly limit the variability of judges in their final judgements. The use of guidelines, for example, has been shown to be effective in the medical field in reducing variability in clinical diagnoses (Kahneman et al., 2021). It has also been demonstrated in the field of education with the use of assessment rubrics (Marin-Garcia and Santandreu-Mascarell, 2015; Rezaei and Lovorn, 2010). It would be expected in the field of scientific assessment they could be very useful as well. Basically, the process of evaluating scientific work is not so different from evaluating educational work (and, if we confine ourselves to the university system, the same people often coincide).

In our opinion, another problem in peer review is reviewers are frequently asked to take a position on whether the manuscript should be published (WPOM includes this, see Table 1). It is quite common for this to appear in the instructions which journals provide to reviewers and/or in the form they must complete. In this situation, whenever an overall assessment is requested, what usually happens (even in the case of expert reviewers) is the judgement process is initiated by making an overall assessment which frequently incorporates an inclination towards a particular conclusion (accepting or rejecting the article). We know as humans we are prone to decide first and then go on to develop arguments to support our initial decision (Ariely, 2008). In this process we tend to selectively gather and interpret arguments in a way that supports the decision already made. This is a well-known and widespread effect (more than is widely believed) which is due to different psychological biases. In short, we form consistent impressions quickly and are slow to change them (Kahneman, 2012). This effect would only be positive if the conclusions are correct, but when the initial assessment is wrong, in whole or in part, the tendency to maintain it in the face of contradictory evidence is highly persistent.

To correct this, at least in part, WPOM (an idea that can apply to other scientific journals) should not have required reviewers to indicate whether the manuscript was publishable or not, or even to make an overall judgement, because we would have been in the same situation. We propose adapting to peer review processes Kahneman et al. (2021) call complex judgement structuring, which consists of applying three principles: decomposition, independence, and delayed holistic judgement. The first, decomposition, is already being applied by journals by asking reviewers to evaluate the manuscript based on a series of differentiated criteria (contribution to knowledge, methodological adequacy, applied relevance, etc.). It should be stressed that the assessment of these criteria should be based on objective information as far as possible. Consequently, reviewers should avoid making general statements about the manuscript they are reviewing that are not supported by facts. In short, they should substantiate the suggestions they make with concrete data and information. For example, if a manuscript has problems with the comprehensibility of the text, they should specify exactly where these problems are to be found.

The second principle, independence, requires performing information collection on each assessment of a criterion in isolation. This does not mean each reviewer acts separately from the other reviewers, as is normally the case, a practice which should continue as it is an effective practice to reduce noise (worth highlighting in the context of this article as it is a practice scientific evaluation has incorporated with success for a many years). Rather, it means emphasising during the process each of the criteria by which the quality of a manuscript is judged must be assessed separately, otherwise each assessment of one criterion influences the others (even if they are unrelated), resulting in each assessment being influenced by previous ones thereby adding significant noise to the whole process. For example, the message should be stressed that it is normal a manuscript can score very well on one criterion and very poorly on the next. Or if the form is completed online, it needs to be designed to prevent people from being able to move on to the next question if they have not answered the previous one, including warnings if they want to go backwards. Ideally, in order to move towards independent judgements, there would probably be reviewers qualified to assess a single criterion separately, specialists in judging a single criterion which they apply to different manuscripts, so for a single manuscript we would have the expert judgement of a set of reviewers without any linkage in their assessments (this would have the added advantage of having to select the assessment criteria carefully, which, if this procedure were followed, could not be very numerous). It may seem this would make peer review processes even more difficult, but we are convinced the opposite is true. A reviewer would only evaluate their criteria in each review, increasing their specialisation and thus reducing their investment in time, effort, etc. This would help to form a stable body of reviewers for journals who, moreover, would find it easy to recognise their vital contribution as outstanding specialists.

The above procedure goes hand in hand, as previously mentioned, with the fact the reviewers will not have to make an overall judgement, this task is reserved exclusively for the editor(s) (i.e., delayed holistic judgement). The editor (although it would be even better if it were a committee of several editors acting independently of each other) will be the one who will make a final judgement from all the information gathered in the review process, as they are really the one who can make an overall assessment of the quality of the reviewed manuscript. The problem is that some web platforms where journals are hosted (OJS, for example, in the case of WPOM) are predefined and, in principle, do not allow the editors of each journal to configure them differently (e.g., pop-up windows, drop-down labels, etc.) as they are the same for all hosted journals. Consequently, some of the changes we propose would have to be undertaken on the web platforms themselves and by seeking consensus with journal editors. This undoubtedly makes it

difficult to introduce some of the proposed changes, but the advantages achieved could outweigh the costs if we want to move towards a less noisy process in scientific evaluation.

This article offers several practical ideas which, in addition to those already in place, would reduce the noise level in peer review processes, e.g., using guidelines, using appropriate and agreed terminology, or adopting the method of structuring complex judgements. It is true noise reduction strategies (decision hygiene) can sometimes be costly (the first step is overcoming the phenomenon of resistance to change). However, often their costs are merely an excuse (Kahneman et al., 2021). In short, the benefits of noise reduction must be compared with its costs, and a decision must be made whether it pays off. In the case of the evaluation of scientific articles, it seems this may be the case, as the proposed measures appear affordable and the expected benefit is likely great because it would improve the decision-making process, make the process fairer, and possibly lead to better science. In conclusion, the first step to resolving the noise level in peer review processes is to recognise the problem exists and that it is possible to measure the noise. This article represents a first approximation of this.

Regarding limitations in our study, we have not addressed the interesting question of how to detect and eliminate bias in peer review because it requires a different treatment than the one presented here (the concepts of bias and noise should not be confused). On the other hand, if all reviewers were able to evaluate more than one manuscript, we could have analysed other aspects of noise, such as level noise which referred, in our case, to the variation among reviewers in their willingness to give severe assessments of the quality of the manuscripts considered (it was known not all referees were equally harsh). Or pattern noise, a more complex aspect of noise analysis which refers to referees not being equally harsh in all their judgements of the manuscripts under their review. They are harsher than their personal average with some manuscripts and more lenient with others, reflecting a complex pattern in the attitudes of judges towards particular cases (the statistical term for pattern noise is referee × case interaction). Both types of noise, level noise and pattern noise, would provide the measurement of total system noise and allow for a more complete noise audit.

Research could be designed with the aim of estimating the level of total noise in peer review processes by developing the necessary material which should include quantitative measurements of the variables considered (which could answer some of the questions posed by relevant contributors in the field of peer review, e.g., Weller, 2001). In this study, which we considered as a pilot, we chose to evaluate a manuscript previously accepted for publication and before this decision was contaminated by the results of the audit. In the future, the analysis could be extended to look at noise in submissions which have been rejected by an editor (or editorial team). Then we could investigate not only the noise of the reviewers, but also the noise of the journal editors. If, instead of choosing one pair of reviewers, a rejected manuscript had another pair of reviewers, would the final editorial decision have been the same? One could even assess the noise level in the submission of manuscripts which are rejected directly by the editor and do not advance to peer review because the editor considers them to be of inadequate scientific quality. How likely would a manuscript receive a positive assessment by the reviewers of that journal? Finally, there is no reference level or cut-off value which can be set as the limit beyond which noise should be considered unacceptable in a peer review process. It is certainly tempting to import cut-off values derived from other fields (Belur et al., 2021; LeBreton and Senter, 2008; Voskuijl and Van Sliedregt, 2002). But the establishment of such cut-off values requires adaptation to the context in which they are applied and, perhaps,

also an analysis of the levels of inter-rater agreement or reliability in scientific journals (Benda and Engels, 2011; Bornmann, 2011).

Conflict of interests

Both authors have collaborated in a similar way throughout the process of drafting this article.

Spanish version

Una auditoría del ruido en la evaluación por pares de un artículo científico. El caso de la revista WPOM

Abstract

El presente estudio pretende ser uno de los primeros en analizar el nivel de ruido en los procesos de revisión por pares de los artículos científicos. Se entiende por ruido la variabilidad no deseada en los juicios que emiten los profesionales sobre un mismo tema o asunto. Nos referimos a juicios evaluativos en los que se espera que los expertos/as tiendan a coincidir. Es lo que sucede cuando se pretende juzgar la calidad de un trabajo científico. Para medir el ruido, la única información que se precisa es disponer de varios juicios emitidos por diferentes personas sobre un mismo caso para analizar su dispersión (lo que Kahneman y colaboradores denominan como "auditoría del ruido"). Este fue el procedimiento que seguimos en esta investigación. Les pedimos a un conjunto de revisores/as de la revista WPOM (Working Papers on Operations Management) que revisaran un mismo manuscrito que ya había sido aceptado previamente para su publicación en esta revista aunque ellos lo desconocían. Los resultados indicaron que en el supuesto de usar dos personas evaluadoras, la probabilidad de que ese manuscrito no hubiese sido finalmente publicado sería cercana al 8% mientras que la probabilidad de que hubiera tenido un futuro incierto sería del 40% (una opinión favorable y otra desfavorable o ambas que plantean cambios sustanciales). En el caso de emplear solo a un revisor/a, en el 25% de las ocasiones el trabajo auditado hubiera tenido muchos problemas para su publicación. La gran ventaja de medir el ruido es que una vez medido es posible por lo general reducirlo. Este artículo termina exponiendo algunas de las medidas que se pueden poner en marcha por parte de las revistas científicas para mejorar sus procesos de revisión por pares.

Keywords: *evaluación por pares; evaluación de revistas científicas; evaluación de la investigación; toma de decisiones; ruido en las decisiones; emitir juicios*

Introducción

Desde la publicación del libro de Kahneman, Siboni y Sunstein en 2021 titulado "Ruido: Un fallo en el juicio humano" y se su anterior precuela (Kahneman et al., 2016) un nuevo tópico ha aparecido en escena con un impacto importante. A principios de 2023 más de 65 artículos científicos indexados en SCOPUS habían citado el trabajo de Kahneman et al. (2016), lo que sitúa este artículo en el percentil 1 de los artículos más citados (por año y área). Sin embargo, por lo que respecta a la investigación científica y, en concreto a los procesos de revisión por pares, aunque mucho es lo que se ha publicado hasta el momento (Bornmann, 2011; Weller, 2001), no lo ha sido siguiendo esta perspectiva basada en analizar el nivel de ruido. Este artículo pretende ser una excepción a la situación anterior. En realidad, creemos que es la primera vez que una revista científica se embarca en una auditoria para evaluar el nivel de ruido existente en sus procesos de revisión por pares.

Se entiende por ruido la "*variabilidad no deseada* en juicios sobre un mismo problema" (Kahneman et al., 2021, p. 19, el subrayado es nuestro). La variabilidad en los juicios no siempre es indeseada, por ejemplo, cuando se trata de cuestiones de opinión, preferencia o de gusto. Pero no hay que confundir gusto personal y juicio profesional. Una cosa es que dos personas, incluso dos especialistas, puedan diferir en qué cuadro de pintura les gusta más, y otra cosa bien distinta es establecer su precio al tasarlo. La palabra «juicio» se utiliza sobre todo cuando se considera que la gente debe estar de acuerdo (Kahneman et al., 2021). Las cuestiones opinión o de gusto difieren de las cuestiones de juicio en que las diferencias son aceptables, incluso deseables. Pero en las cuestiones de juicio, lo que esperamos es que los expertos/as tiendan a coincidir y las diferencias se reduzcan al mínimo. Si usted padece una enfermedad, por ejemplo, no le agradará recibir dos diagnósticos muy diferentes, entre otros motivos porque pensará con razón, que al menos uno de los dos necesariamente ha de estar equivocado. Si están evaluando su desempeño en su empresa, no entenderá que dos profesionales de los recursos humanos emitan juicios muy dispares sobre cuál es su verdadero nivel de rendimiento real. Y así podríamos continuar poniendo innumerables ejemplos en los que una variabilidad excesiva en los juicios no se consideraría aceptable por nadie (p. ej., a la hora de determinar el grado de minusvalía de una persona, o la sentencia penal que debe recibir un acusado, o las calificaciones en las pruebas de acceso a la universidad de sus hijos/as, y así un amplio etcétera).

Es cierto que en la publicación científica se asume que un mismo manuscrito puede ser valorado de forma muy diferente por distintos evaluadores (Benda y Engels, 2011; Theoharakis et al., 2007). No son pocas las revistas que señalan en sus instrucciones para los autores, que remitirán los manuscritos a dos revisores anónimos y que, en caso de que no haya acuerdo entre ellos/ellas, lo remitirán a un tercero teniendo siempre la última palabra el editor/a de la revista. Se trata de una práctica común en todas las disciplinas científicas, no solo en ciencias sociales. Esto no significa otra cosa que asumir que, en ciencia, no existe un criterio único que permita validar la calidad con que ha sido elaborado un trabajo, o de que, aunque exista cierto consenso en estos criterios, no necesariamente son aplicados de igual forma por los propios investigadores (Bedeian, 2004). Como afirman Kahneman et al. (2021), aceptar un cierto nivel de variabilidad es asumible cuando el que decide es el ser humano. Pero no debemos olvidar que estamos hablando de juzgar la calidad de un trabajo científico, no es una mera cuestión de opinión. Es una decisión similar a la que toma un juez o jueza en sus sentencias (no se espera que, ante un mismo delito, dos jueces puedan emitir sentencias muy distintas). Entonces nos preguntamos ¿es justo que un mismo manuscrito enviado a una revista pueda ser valorado de forma tan diferente?: desde su rechazo directo hasta su aceptación con mínimos cambios, pasando por las valoraciones intermedias que sugieren modificaciones con un grado de variabilidad muy notable.

En resumen, podemos asumir que los juicios evaluativos implican una expectativa de desacuerdo limitado. Como afirman Kahneman et al. (2021) los juicios evaluativos dependen, en parte, de los valores y preferencias de quienes los hacen, pero no son meras cuestiones de gusto o de opinión. Una gran variabilidad en los juicios sobre un mismo caso siempre será indeseada. Siempre que la persona que ha de emitir un juicio es seleccionada al azar, o casi, dentro de un conjunto de personas igualmente cualificadas (como es el caso de la revisión por pares en ciencia), el ruido es un problema. Todo lo anterior implica asumir por lógica un determinado grado de azar en la elección de los artículos que finalmente son publicados. Puede hasta llegar a convertirse en una lotería (dependiendo del revisor/a que te toque, el artículo verá la luz o no). Partimos del supuesto erróneo de que otro experto/a llegaría a un juicio similar ante un mismo

estímulo (por ejemplo, un manuscrito científico), sin embargo, lo frecuente es que no sea así (Ernst et al., 1993). En diferentes estudios realizados en distintos ámbitos (sentencias judiciales, agentes de seguros, etc.), Kahneman et al. (2021) han encontrado que el ruido, en promedio, es cinco veces mayor de lo que se pudiera pensar inicialmente. En el caso específico de la revisión por pares, el acuerdo entre jueces oscila entre 0.20 y 0.40, lo que se corresponde con un nivel bajo de acuerdo (Bornmann, 2011).

Nosotros proponemos hacer una primera aproximación a este asunto desde una óptica diferente. La evaluación por pares es un juicio evaluativo que, por su naturaleza, no es verificable. No obstante, "todo lo que necesitamos para medir el ruido son múltiples juicios sobre el mismo problema. No necesitamos conocer un valor real" (Kahneman et al., 2021, p. 27). La única información que se precisa consiste en disponer de varios juicios emitidos por diferentes personas sobre un mismo caso y analizar su dispersión, para comprobar qué valores alcanza. "Auditoría del ruido" es el nombre que Kahneman et al. (2021) utilizan para referirse a este procedimiento. ¿Qué necesitamos, pues, para comenzar? Que diferentes evaluadores realicen juicios sobre los mismos casos, en este artículo que presentamos, que diferentes *referees* revisen un mismo manuscrito que ha sido enviado a una revista científica para su publicación, para analizar a continuación el grado de variabilidad en sus juicios. Además, una vez medido el ruido, es posible por lo general reducirlo. Es lo que Kahneman et al. (2021) denominan como "higiene de las decisiones", concepto que hace referencia al conjunto de estrategias y técnicas que se pueden emplear para reducir el ruido.

Método

Para este trabajo se eligió el último manuscrito aceptado en la revista *Working Papers on Operations Management* (WPOM), pero no publicado, en el momento de escribir la versión inicial de este artículo (inicio de 2022). Ese manuscrito completó el proceso normal de revisión de la revista (dos revisores/as evaluaron el trabajo en un formato doble ciego y el editor jefe tomó la decisión editorial de aceptarlo), y todo ello fue antes de que se le planteara al editor jefe (autor 2) la colaboración en esta investigación sobre auditoría del ruido.

Se solicitó por correo electrónico a los autores del manuscrito su consentimiento informado para que su trabajo se incluyera en esta investigación (ver anexo 1). Recibimos la conformidad de todos los autores el 16 de febrero de 2022.

El editor jefe de la revista seleccionó a 24 revisores/as de la revista entre las personas más activas y con mejor cumplimiento de plazo en las revisiones. Todas ellas eran doctoras con años de experiencia, posiciones estables en la academia y varios artículos publicados en revistas de reconocido prestigio internacional. Se utilizó el correo de invitación estándar de la revista, sin añadir ninguna información adicional sobre esta auditoría, y se les dio el plazo habitual (una semana para aceptar o rechazar el encargo y 4 semanas para completar la revisión). Su participación, como es habitual en la revista, es confidencial y todos los datos se han anonimizado antes de ser procesados durante la investigación. Por lo tanto, el proceso de revisión fue "ciego" durante esta auditoría del ruido.

Finalmente 18 personas decidieron aceptar la solicitud de revisión. Todas ellas usaron la plantilla de revisión oficial de la revista para protocolos de revisión sistemática de la literatura (el tipo de artículo usado para esta prueba de auditoría del ruido). Los ítems de esta plantilla de revisión están basados en la declaración PRISMA (2009), entre otras.

Los autores del manuscrito seleccionado para esta auditoría recibieron una copia de los comentarios de las 18 personas revisoras extra. Estos comentarios fueron considerados muy valiosos por estos autores, que aprovecharon los comentarios recibidos para mejorar su investigación, que ha sido finalmente publicada (Álvarez y Maheut, 2022). Por lo que, podemos afirmar, que el trabajo de revisión extra realizado ha sido aprovechado para la consecución de una mejor ciencia publicada.

Habitualmente, la fiabilidad entre jueces se mide con kappa de Kripendorf y otras medidas similares (Krippendorff, 2011; LeBreton y Senter, 2008). Cuanto mayor sea el estadístico kappa, menor será el ruido (un valor kappa de 1 refleja un acuerdo total). En esta auditoría, al tener solo un objeto a evaluar por 18 jueces, no es posible usar estas medidas de fiabilidad entre jueces (ver para una revisión de estas, por ejemplo, Benda y Engels, 2011; o Weller, 2001), por lo que hemos analizado la concordancia (grado de acuerdo) entre las personas que participaron en el proceso de revisión y realizado un análisis de probabilidades, estimando el porcentaje de clasificaciones iguales frente al número de alternativas de clasificación posibles.

Resultados

Comenzamos la descripción de los resultados de nuestro estudio con la valoración global que realizan los revisores/as del manuscrito que les pedimos evaluar.

Atendiendo a la Tabla 1, hay que tener presente que, al existir 4 alternativas de respuesta, el máximo nivel de ruido se obtendría si cada alternativa fuera seleccionada un 25% de los casos, máxima dispersión. El mínimo nivel de ruido se obtendría cuando solo una de las cuatro alternativas fuese seleccionada en el 100% de los casos, máximo nivel de concordancia.

Tabla 1. Recomendación final de las personas revisoras acerca del manuscrito enviado

	Frecuencia	Porcentaje
Rechazar envío	2	11.1
Cambios mayores	3	16.7
Cambios menores	10	55.5
Aceptar envío	3	16.7
Total	18	100

En la revisión de artículos científicos, dado que el paso a revisión es tras una evaluación previa de una persona editora científica de la revista, la opción de respuesta más probable es que los revisores/as soliciten cambios a los autores. Estos cambios pueden ser menores o pueden suponer grandes modificaciones en el manuscrito revisado (los términos comúnmente utilizados en inglés son *minor revision* y *major revision*). En WPOM se sigue la misma distinción (cambios menores y mayores), además del rechazo o la aceptación directa del manuscrito. El proceso de decisión habitual en WPOM es el siguiente: a) si todas las opiniones de revisores/as son favorables (aceptar tal cual o cambios menores), el trabajo es aceptado para su publicación y la versión definitiva será solo revisada por el editor/a que esté a cargo del envío; b) cuando todas las opiniones son negativas (rechazar el envío), el editor jefe valora las críticas realizadas por las personas que revisan y, salvo que las opiniones de las personas revisoras no estén debidamente justificadas, el artículo será rechazado; c) en cualquiera de los otros casos, se ofrece a las autoras/es la

opción de pasar a una nueva ronda de revisión, si bien, en el supuesto de haber recibido críticas mayores por parte de alguno de los revisores/as, las posibilidades de publicación se reducen drásticamente.

Dado que el manuscrito usado en esta auditoría ya había sido aceptado (como se ha descrito en el apartado de Procedimiento), en caso de ausencia o bajo nivel de ruido, la gran mayoría de los revisores/as hubieran coincidido en una misma alternativa alineada con la decisión editorial. Es decir, las recomendaciones deberían haber sido entre aceptarlo o introducir cambios menores. Sin embargo, como se puede apreciar en la Tabla 1, la opción con el porcentaje más alto (cambios menores) se sitúa muy poco por encima del 50%. En total 13 de las 18 revisiones (el 72%) han coincidido con la decisión editorial ya tomada. El 28% restante se han inclinado por una valoración claramente diferente (sugerir cambios mayores o recomendar su rechazo directo). Lo que significa que en 1 de cada 4 ocasiones este manuscrito hubiera tenido muchos problemas para su publicación o no se hubiera publicado directamente, y todo ello teniendo en cuenta que se trata de un mismo manuscrito que ya había seguido el proceso de revisión por pares habitual en cualquier revista científica.

Análisis de probabilidades

Cada persona revisora puede elegir entre 4 recomendaciones (rechazo, cambios mayores, cambios menores y aceptación). El número de combinaciones con repetición de n elementos tomados de k en k se puede calcular con la fórmula:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Considerando que haya dos personas revisoras (situación habitual en el proceso de revisión estándar de una revista) y que el orden de las recomendaciones no es relevante (es decir, ser revisor 1 o 2 no da más peso a la recomendación) tenemos que, en nuestro caso, el número de combinaciones de 4 elementos con repetición tomados de dos en dos es 10, siendo las otras 6 opciones una situación análoga a alguna de esas 10. La Tabla 2 recoge las opciones de combinación y las distintas situaciones en las que puede quedar un manuscrito tras la revisión por pares en WPOM (en otras revistas podría ser diferente).

Tabla 2. Distintas combinaciones posibles tras un proceso de revisión por pares con 4 alternativas de valoración

	Revisor/a 2→	Rechazar envío	Cambios mayores	Cambios menores	Aceptar envío
Revisor/a 1↓					
Rechazar envío		Rechazar	Difícil de publicar	Indeciso	Indeciso
Cambios mayores		Difícil de publicar	Difícil de publicar	Indeciso	Indeciso
Cambios menores		Indeciso	Indeciso	Publicar	Publicar
Aceptar envío		Indeciso	Indeciso	Publicar	Publicar

Tomando como referencia el uso de dos personas como revisoras (la situación más habitual en cualquier revista), si las recomendaciones se distribuyeran de manera uniforme (misma probabilidad para cada una de las opciones, igual a 0.25), cada una de las casillas de la Tabla 2 tendrían un 6.25% de probabilidad (0.25*0.25). En ese caso, teniendo en cuenta el proceso de decisión habitual en WPOM (que se ha descrito en el apartado anterior), la probabilidad de un rechazo directo sería de un 6.25%, una nueva ronda de revisión difícil de publicar sería equivalente al 18.75%, un 50% de los manuscritos tendría un futuro incierto y el 25% restante se publicaría con seguridad (bien directamente o bien tras unos cambios sencillos y rápidos en el manuscrito original). En el caso de ausencia de ruido, es decir, si siempre coincidieran los dos revisores/as, sólo serían posibles las celdas de la diagonal (cada una con un 25% de probabilidad), de modo que la probabilidad de rechazo sería de un 25%, la de una nueva ronda de revisión difícil de publicar también del 25%, y un 50% de los manuscritos serían publicados (ningún artículo tendría un futuro indeciso).

No obstante, los artículos que entran en un proceso de revisión han pasado por el filtro previo del editor/a jefe y no parece muy razonable que las cuatro opciones de recomendación sean equiprobables. Y si, como es nuestro caso, el manuscrito ya estaba aceptado para ser publicado, mucho menos. Por lo tanto, lo esperable sería que los revisores/as que han participado en nuestro estudio, en condiciones de ausencia o bajo ruido, hubieran tendido a coincidir determinando que su valoración fuese publicarlo.

Extendiendo el análisis de la Tabla 2 a la situación de esta auditoría que cuenta con 18 personas como revisoras, se pueden dar 153 parejas de revisores/as distintas. Además, en lugar de suponer una distribución uniforme de las evaluaciones de los revisores/as, usamos las frecuencias reales de las decisiones tomadas. En este caso, la probabilidad de un rechazo directo es de un 1.2% (probabilidad de que, en el total de las 153 parejas, dos personas coincidan en rechazar el manuscrito), la probabilidad de que se proponga una nueva ronda de revisión que difícilmente acabaría en publicación sería del 6.5%, un 40.1% de los manuscritos tendrían un futuro incierto y el 52.1% restante se publicaría con bastante seguridad. Es decir, con un 50% de probabilidad, las posibles combinaciones al azar de las personas que han participado en esta auditoría del ruido se alinearían con la decisión editorial de aceptar el manuscrito, lo que implica en consecuencia que el otro 50% restante se podría considerar ruido.

Análisis detallado de la evaluación por pares

Los resultados anteriores están referidos a la decisión final que se puede tomar acerca de un artículo que se revisa por pares. Por otro lado, la revista WPOM utiliza una plantilla con diferentes ítems para evaluar cada uno de los manuscritos adaptada según el tipo de trabajo que recibe (para esta auditoría se trató de una revisión sistemática de la literatura). En la Tabla 3 mostramos el desglose de las puntuaciones cuantitativas asignadas a cada uno de los ítems por parte de las personas que revisaron ese manuscrito.

Para todos los ítems analizados en esta Tabla 3, el mayor nivel de ruido se hubiera obtenido si el 33% de las respuestas de los revisores/as se hubieran repartido por igual entre las tres alternativas posibles en cada variable, a excepción de los ítems 3, 11 y 13, en el que el máximo porcentaje de dispersión sería del 25% al existir 4 alternativas de respuesta. Se observa que el máximo nivel de acuerdo se alcanza en la variable 10 y que, igualmente es muy elevado en las variables 7, 8 y 9. Por el contrario, en el resto de los ítems existe un nivel de ruido variable. Es relativamente bajo en la variable 1, aumenta un poco en los ítems 2, 3, 4, y 5, mientras que es ya más alto para las variables 6, 12, 13 y 14. Asimismo, uno podría

estar tentado a pensar que, en esta Tabla 3, el revisor que indica la opción "not at all" en la variable 2 es el mismo/a que elige idéntica alternativa en las variables 3 y 4, pero no es así, todos ellos son personas diferentes. Por señalar otro ejemplo, si comparamos lo que sucede con las variables 7, 8 y 9, no coincide ninguno de los revisores/as que se desmarcan de la abrumadora mayoría. En suma, no siempre existe una pauta unívoca en la elección que hacen los revisores de este manuscrito, quien elige una opción en un ítem frecuentemente la cambia en otro por lo que, en consecuencia, las respuestas dadas se combinan de diferentes modos.

En el primer ítem, ante la pregunta de si el manuscrito contiene todas las partes de un protocolo para la revisión de la literatura, una amplia mayoría de los revisores/as opta por reconocer que se cumple suficientemente con este criterio y nadie considera que suceda lo opuesto. De todos modos, casi un tercio estima que el criterio se cumple marginalmente y necesita ser mejorado. En cuanto a los comentarios emitidos por los revisores/as en este punto encontramos (ver Tabla 4), desde aquellos que están francamente a favor (por ejemplo, "El artículo está bien estructurado y sigue la estructura habitual de este tipo de artículos" o "El protocolo de revisión sistemática de la literatura es riguroso y se sigue conforme los pasos establecidos en trabajos de referencia") hasta aquellos otros que muestran justo lo contrario (p. ej. "La estructura del artículo no se ajusta perfectamente a un protocolo de revisión de la literatura" o "La revisión de la literatura solo contiene los análisis descriptivos básicos y más típicos"). Por supuesto, pasando por comentarios de naturaleza intermedia ("Aunque los apartados exigidos están, se sugiere mencionar también cómo ha realizado el proceso de codificación posterior y la inclusión de un plan de trabajo tentativo").

En relación al ítem 2 acerca de si el manuscrito es comprensible para alguien que no es un experto/a, los resultados son muy similares a la variable anterior, tan solo que en este caso en una revisión se considera que de ningún modo se cumple con este criterio. Igual sucede con los comentarios, siendo muy positivos en algunos casos ("Se entiende perfectamente. Tiene una fácil lectura y permite tener un conocimiento adecuado de la técnica y sus objetivos" o "El trabajo está escrito correctamente"), muy negativos en otros ("Aunque el contenido del artículo es muy básico, tiene serios problemas de redacción, lo que no facilita su lectura"), o que muestran matices varios ("Aunque el texto se entiende por alguien que no sea experto, presenta errores ortográficos, de formato y de puntuación, que deben ser corregidos").

Podríamos continuar describiendo uno por uno el resto de las variables, pero para hacer más ágil la lectura, hemos preferido incluir la Tabla 4 que recoge ejemplos de comentarios que realizan las personas que realizaron la revisión de ese mismo manuscrito. A la hora de elaborar esta Tabla 4, se ha llevado especial atención de incluir comentarios de todos los revisores/as siempre que fuese posible, porque no todos incluían comentarios a cada variable. Además, se ha puesto un especial esfuerzo en introducir comentarios de todas las revisiones, no únicamente de las que mantenían una opinión francamente favorable o desfavorable del manuscrito en su conjunto. Asimismo, como se puede apreciar, se han combinado comentarios positivos y negativos de diferentes revisores/as que parecen contraponerse entre sí (en una revisión se resalta algún aspecto como positivo y otro como negativo, mientras que en otra revisión se hace justo lo contrario).

En esta Tabla 4 se puede apreciar que existen desde sugerencias que no se enmarcan en la misma dirección, aunque se les esté preguntando sobre un mismo aspecto, hasta comentarios manifiestamente contrapuestos entre sí. No es simplemente que los revisores/as comenten diferentes aspectos del manuscrito,

como se ha argumentado para justificar la baja fiabilidad entre jueces (Fiske y Fogg, 1990). Entonces ¿en qué sentido deberían mejorar su trabajo los autores del manuscrito revisado?, ¿cómo podrían integrar sugerencias que en el fondo están planteando ideas contradictorias? Parece claro que, dependiendo del azar (no hay que olvidar que lo habitual es enviar a dos expertos/as anónimos la revisión de un manuscrito, escogidos de una "lista" de potenciales revisores/as), el artículo final que se publicaría podría llegar a ser muy distinto (o no ser publicado). Como Bornmann (2011) señala, son muy pocos los estudios que han investigado las razones reales detrás del desacuerdo de los revisores/as (por ejemplo, realizando análisis comparativos del contenido de las hojas de revisión), nuestro trabajo supone una nueva aportación en esta línea.

Tabla 4. Comentarios literales (traducidos del inglés cuando ha sido preciso) realizados por las personas revisoras del manuscrito

Número	Ítem	Ejemplos de comentarios positivos	Ejemplos de comentarios negativos
1 SLRjam01	Does it contain all the parts of a literature review protocol?	El artículo está bien estructurado y sigue la estructura habitual de este tipo de artículos (revisión 1)	La estructura del artículo no se ajusta perfectamente a un protocolo de revisión de la literatura (revisión 5)
2 SLRjam02	Is it understandable by someone who is not an expert?	El trabajo está escrito correctamente (revisión 5)	Aunque el contenido del artículo es muy básico, tiene serios problemas de redacción, lo que no facilita su lectura (revisión 11)
3 SLRjam03	Are all the "variables" properly defined?	Las variables utilizadas para el análisis son adecuadas (revisión 11)	El artículo no define las variables (revisión 8)
4 PRISMA03	Does it describe the rationale for the review in the context of what is already known?	Los autores añaden un apartado específico para analizar el estado actual de las revisiones de literatura sobre el tema tratado (revisión 5)	Los autores deberían esforzarse más en explicar por qué es necesaria esta revisión de la literatura y cómo estos resultados podrían ser útiles para futuras investigaciones (revisión 4)
5 PRISMA04	Does it provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design?	Los autores plantean algunas preguntas de investigación y se hace explícito el contexto para la revisión de la literatura (revisión 4)	Entiendo que no se da una explicación detallada de todos estos elementos (revisión 2)
6 SLRTemp2d	If extending previous research on the topic, does it explain why a new study is needed?	Los autores muestran tres artículos basados en revisiones de literatura y los explican (revisión 8)	Falta explicitar qué va a aportar este estudio a los ya existentes (revisión 10)
7 SLRTemp3a	Specify and justify basic strategy: manual search, automated search, or mixed	Entiendo que sí se explica este punto en el artículo (revisión 2)	Se solicita aclaración sobre alguno de los pasos en documento adjunto (revisión 9)
8 PRISMA06 SLRTemp4a-b	Identify the inclusion criteria for primary studies, identify the exclusion criteria	Los criterios de inclusión y exclusión son explícitos y parecen razonables (revisión 4)	Sólo se listan, se debería haber dado más detalle (revisión 2)

9 PRISMA07 SLRTemp3c	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched	El artículo describe las fuentes de información usadas para la búsqueda de información (revisión 11)	Ver comentarios de mejora en archivo adjunto (revisión 3) -en ese adjunto se incluían numerosos comentarios-
10 PRISMA08 SLRTemp3b	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	Sí, se eliminan los artículos que están repetidos en las bases de datos utilizadas (revisión 5)	Ver comentarios de mejora en archivo adjunto (revisión 3) -en ese adjunto se incluían numerosos comentarios-
11 SLRTemp3d	For manual searches, identify the journals and conferences to be searched	Sí, se especifican (revisión 2)	No aplica (revisión 11)
12 SLRTemp3e	Specify the time period to be covered by the review and any reasons for your choice	El período de tiempo fue de 2012 a 2020 (revisión 4)	No se especifica la justificación del horizonte temporal considerado para la revisión de la literatura (revisión 14)
13 SLRTemp3f	Identify any ancillary search procedures, e.g., asking leading researchers or research groups, or accessing their web sites; or checking reference lists of primary studies	Puede ser interesante analizar la bibliografía de los 40 artículos seleccionados, con el fin de confirmar la estrategia de análisis e incorporar artículos importantes (revisión 7)	No realizan una búsqueda de bola de nieve (revisión 16)
14 SLRTemp3g	Specify how the search process is to be evaluated, (e.g., against a known subset of papers, or against the results from a previous systematic review)	Se comparan los resultados obtenidos con revisiones previas (revisión 9)	No se compara ni discute con los resultados de revisiones sistemáticas similares (revisión 14)

Discusión

En esta investigación, optamos por enviar a las personas revisoras un manuscrito que ya había sido previamente aceptado por el editor de la revista tras un proceso de revisión por pares previo, mediante el cual se incorporaron diferentes modificaciones a la versión inicial del manuscrito. Este hecho, debería haber reducido en gran medida el ruido en la evaluación del mismo, ya que se trataba de un artículo mejorado y aceptado para ser publicado tal cual. Y así ha sido en parte. De todos modos, los resultados indican que el ruido existe y que no es despreciable. En principio, los revisores/as que se distancian sensiblemente de la aceptación del artículo provocan ruido (en este caso, los rechazos y revisiones mayores, suponen un porcentaje superior al 25%). Es decir, uno de cada cuatro revisores/as no hubiera sugerido publicar este manuscrito que ya había seguido un proceso de revisión estándar y que había sido aprobado para su publicación. Teniendo en cuenta que "el nivel de acuerdo entre los revisores/as generalmente no es mucho más alto que el que se esperaría que ocurriera basándose únicamente en el azar" (Benda y Engels, 2011, p. 169), el proceso de revisión por pares seguido por WPOM ha reducido ese porcentaje (asumiendo que la respuesta correcta es aceptar el artículo para su publicación).

Consideramos que un mismo trabajo científico, un manuscrito, no puede ser malo, bueno y regular al mismo tiempo. De igual modo, es injusto que se trate de forma diferente a sus autores, que pueden recibir una valoración favorable o desfavorable dependiendo del revisor/a que les toca, cuando resulta que ambos revisores/as están juzgando exactamente el mismo manuscrito. "Un sistema en el que los juicios profesio-

nales son a todas luces incoherentes pierde credibilidad" (Kahneman et al., 2021, p. 167). Lo importante en los procesos de revisión, es evaluar el trabajo de investigación que se presenta. Sin embargo, nuestro estudio pone claramente de manifiesto que las personas evaluadoras juegan un papel fundamental. Es algo que ya sabíamos, lo relevante ahora es que se aportan más pruebas de que su función es determinante a la hora de juzgar la calidad de una aportación, y probablemente su resultado, hasta el punto de que en algunos casos su función transciende la propia aportación dejando a esta en segundo plano. La aceptación o no de un artículo para su publicación por parte de una revista científica, puede correr el riesgo de convertirse en ocasiones en una lotería, en una cuestión de suerte (Bedeian, 2004). Como afirman Kahneman et al. (2021, p. 27) "en cualquier sistema en el que se supone que los juzgadores son intercambiables y se asignan de forma casi aleatoria, los grandes desacuerdos sobre el mismo caso quebrantan las expectativas de justicia y coherencia." Es un problema que afecta de pleno a la objetividad científica.

Comencemos, pues, a diagnosticar algunos problemas en los procesos de revisión por pares y a proponer posibles soluciones. Para empezar, la propia escala de respuesta supone una fuente de ruido. Las personas pueden diferir en sus juicios, no porque en el fondo estén en desacuerdo, sino porque asignan a los valores de la escala diferentes significados. La escala de respuesta utilizada en la Tabla 3 no parece que pueda plantear muchos problemas en este sentido ya que es clara a priori (o, al menos, las consecuencias de elegir una u otra opción no son determinantes). No ocurre lo mismo, en nuestra opinión, con la escala que se emplea para emitir un juicio global acerca de la calidad del manuscrito (ver Tabla 1). En este caso, la diferencia que supone sugerir "cambios menores" (*minor revision*) o "cambios mayores" (*major revision*) a un manuscrito, puede no ser evidente. De hecho, la experiencia común demuestra que, para lo que un revisor/a son cambios menores, otros revisores/as los consideran cambios mayores, y viceversa. Una directriz clara por parte de las revistas de lo que cada una de ellas considera que son cambios menores o mayores, podría ser muy útil en los procesos de revisión porque se reduciría con ello la ambigüedad, ya que se sabe que las escalas ambiguas son ruidosas (Kahneman et al., 2021).

En el caso concreto de WPOM, las personas revisoras reciben estas aclaraciones adicionales, aunque puede que no sean suficientes: a) para cambios menores; " Las revisiones no serán objeto de una nueva ronda de revisiones por pares", y b) para cambios mayores; "Las revisiones estarán sujetas a una nueva ronda de revisiones por pares". En otros casos, las revistas podrían optar por especificar qué entienden por cambios menores, por ejemplo, reservando este tipo de cambios a cuestiones de redacción, formato, estructura, etc. Y concretar que es necesario acometer cambios mayores cuando es preciso modificar aspectos teóricos, metodológicos, de enfoque o que tengan que ver con la discusión o las conclusiones del estudio. Cada revista podría determinar lo que considerara oportuno en las diferentes alternativas de respuesta que ofrece para cada criterio. Como es frecuente que los revisores/as evalúen trabajos de diferentes revistas, sabrían mejor a qué atenerse en cada una de ellas.

Por otra parte, las variaciones en el nivel de ruido pueden deberse a diferentes motivos, no solamente a que las personas revisoras tienen una opinión distinta sobre el grado de cumplimiento de los criterios de revisión establecidos por las revistas (ver Tabla 3 para conocer estos criterios en el caso concreto de WPOM). También puede deberse a que el enunciado de los criterios -ítems- no siempre es claro, pudiendo ser interpretados de diferentes maneras. Los revisores/as pueden dudar acerca de lo que se les está preguntando, dificultando la comprensión de las variables y aumentando consecuentemente el nivel de ruido. O peor, pueden no dudar pero ello no impide que diferentes revisores/as interpreten de distinto modo un mismo criterio porque en el fondo también resulta ambiguo (la ambigüedad puede estar relacio-

nada tanto con los valores de la escala de respuesta a un ítem, como se ha comentado con anterioridad, como con la formulación del ítem). Si este fuera el caso para algunas de las variables, la solución para reducir el nivel de ruido sería clarificar el enunciado de las preguntas que se le hacen a los revisores/as. Existen directrices estándar para la presentación de informes (PRISMA, STROBE, MOOSE, SQUIRE…) que algunas revistas aconsejan que sus revisores/as utilicen (aunque en el estudio realizado por Hirst y Altman, 2012, solo el 35% de las revistas proporcionaban instrucciones online de algún tipo a sus revisores/as). Sea como fuere, si se especificasen mejor estas directrices y se consiguiera reducir su ambigüedad (se podría testear cómo las interpretan los revisores/as más habituales de una revista, lo que supondría disponer de un sistema todavía más robusto), el proceso de revisión por pares ganaría en objetividad.

En base a los resultados de este estudio, la falta de claridad en la formulación de los ítems 3, 11, 13 y 14, podría ser la causa de su mayor nivel de ruido. Ante esta situación, desde WPOM se podría pensar en reformular los enunciados para hacerlos más claros (y eso que esta revista utiliza PRISMA, aunque sigue sin parecer suficiente), añadiendo alguna información adicional muy concreta que precisara el sentido de lo que se pregunta. Porque se sabe que uno de los mayores problemas que provocan ruido es la falta de una terminología adecuada y, sobre todo, consensuada (Kahneman et al., 2021). Su solución pasa por garantizar un marco de referencia común. Por ello, las directrices pueden ser un poderoso mecanismo de reducción tanto del sesgo como del ruido, porque limitan directamente la variabilidad de los jueces en sus juicios finales. El uso de directrices, por ejemplo, ha demostrado su eficacia en el campo de la medicina para reducir la variabilidad de los diagnósticos clínicos (Kahneman et al., 2021). También se ha demostrado en el ámbito educativo con el uso de rúbricas de evaluación (Marin-Garcia y Santandreu-Mascarell, 2015; Rezaei y Lovorn, 2010). Sería de esperar que, en el ámbito de la evaluación científica, también pudieran resultar muy útiles. Pues, en el fondo, el proceso de evaluar trabajos científicos no se diferencia tanto de evaluar trabajos educativos (y, si nos circunscribimos al sistema universitario, coinciden frecuentemente las mismas personas).

En nuestra opinión, otro problema en la revisión por pares es que frecuentemente se solicita a los revisores/as que se posicionen acerca de si el manuscrito debe ser publicado o no (WPOM lo incluye, ver Tabla 1). Es muy común que aparezca en las instrucciones que las revistas proporcionan a los revisores/as y/o en el formulario que tienen que cumplimentar. En esta situación, siempre que se pide hacer una valoración global, lo que sucede habitualmente (aunque se trate de evaluadores expertos/as) es que se inicie el proceso de enjuiciamiento haciendo una valoración de conjunto que frecuentemente incorpora la inclinación hacia una conclusión particular (aceptar o rechazar el artículo). Ahora sabemos que el ser humano, es muy propenso a tomar primero una decisión para pasar a elaborar argumentos con posterioridad que apoyen nuestra decisión inicial (Ariely, 2008). En este proceso tendemos a reunir e interpretar los argumentos de forma selectiva de modo que avalen la decisión ya tomada. Es un efecto muy conocido y extendido (más de lo que se cree) que se debe a diferentes sesgos psicológicos. En definitiva, nos formamos impresiones coherentes con rapidez y somos lentos en cambiarlas (Kahneman, 2012). Este efecto solo sería positivo si las conclusiones son correctas, pero cuando la evaluación inicial es errónea, total o parcialmente, la tendencia a mantenerla frente a las pruebas que la contradicen es sumamente persistente.

Para corregirlo al menos en parte, WPOM (idea que es generalizable al resto de revistas científicas) no debería exigir a los revisores/as que indiquen si el manuscrito es publicable o no, ni tan siquiera que emitan un juicio global porque estaríamos ante la misma situación. Proponemos adaptar a los procesos de revisión por pares lo que Kahneman et al. (2021) denominan “estructuración de juicios complejos” que

consiste en aplicar tres principios: descomposición, independencia y juicio holístico retardado. El primero, la descomposición, ya viene siendo aplicado por las revistas al pedir a los revisores/as que evalúen el manuscrito en base a una serie de criterios diferenciados (contribución al conocimiento, adecuación metodológica, relevancia aplicada…). Hay que subrayar que la evaluación de estos criterios debe basarse en información objetiva en la medida de lo posible. En consecuencia, los revisores/as deberían evitar hacer afirmaciones generales sobre el manuscrito que revisan que no estén apoyadas en hechos. En definitiva, deben argumentar con datos e informaciones concretas las sugerencias que emiten. Por ejemplo, si un manuscrito presenta problemas de comprensión del texto habría que especificar dónde se encuentran exactamente.

El segundo principio, la independencia, requiere que la información sobre cada evaluación de un criterio se recoja de forma aislada. No se refiere a qué cada revisor actúe por separado del resto de revisores/as, como normalmente ocurre, práctica que hay que seguir manteniendo ya que es una de las mejores prácticas para reducir el ruido (por lo que conviene ponerla en valor en el contexto de este artículo, ya que se trata de una práctica que la evaluación científica ha incorporado con gran éxito desde hace mucho tiempo). Más bien supone enfatizar que durante el proceso, ha de avaluarse cada uno de los criterios por los que se juzga la calidad de un manuscrito por separado ya que, en caso contrario, cada evaluación de un criterio influye sobre los demás (aunque no estén relacionados), lo que provoca que cada evaluación esté contaminada por las anteriores y todo el proceso acabe siendo muy ruidoso. Por ejemplo, hay que insistir en el mensaje de que es normal que un manuscrito pueda ser puntuado muy bien en un criterio y muy mal en el siguiente. O si el formulario se rellena online, tiene que diseñarse para impedir que se pueda pasar a la siguiente pregunta si no se ha respondido a la anterior, incluyendo advertencias si se quiere ir hacia atrás. Lo ideal para avanzar en la emisión de juicios independientes, probablemente sería que hubiera revisores/as cualificados en evaluar un único criterio de manera separada, especialistas en juzgar un solo criterio que aplican a diferentes manuscritos, de modo que para un mismo manuscrito tendríamos el juicio experto de un conjunto de revisores/as sin ninguna vinculación en sus valoraciones (tendría la ventaja añadida de que habría que seleccionar muy cuidadosamente los criterios de evaluación que, de seguir este procedimiento, no podrían ser muy numerosos). Puede parecer que esto dificultaría todavía más los procesos de revisión por pares, pero estamos convencidos justamente de lo contrario, porque un revisor/a solo evaluaría su criterio en cada revisión, aumentando su especialización y reduciéndose así su inversión en tiempo, esfuerzo, etc., lo que ayudaría a encontrar un cuerpo de revisores estable por parte de las revistas a los que, por otra parte, resultaría muy fácil reconocerles su inestimable contribución como especialistas destacados.

Todo el procedimiento anterior va unido, como se ha señalado previamente, a que los revisores/as no van a tener que emitir un juicio global sino que esta tarea queda reservada exclusivamente para el editor/es (juicio holístico retardado). El editor/a (aunque sería mejor que fuese un comité de varias personas editoras que actúen de forma independiente los unos de los otros) será quien emitirá una valoración final a partir de toda la información recopilada en el proceso de revisión, pues es realmente quien puede hacerse una composición general de la calidad del manuscrito revisado. El problema es que algunas plataformas web donde se alojan las revistas (OJS, por ejemplo, para el caso de WPOM) vienen predeterminadas y, en principio, no permiten a los editores/as de cada revista configurarlas de modo diferente (ventanas emergentes, etiquetas desplegables, etc.) ya que son iguales para todas las revistas alojadas. En consecuencia, algunos de los cambios que estamos proponiendo tendrían que acometerse desde las propias plataformas

web buscando el consenso con los editores/as de las revistas. Sin duda lo anterior dificulta la introducción de parte de los cambios propuestos, pero las ventajas conseguidas podrían compensar los costes si se quiere avanzar hacia un proceso menos ruidoso en la evaluación científica.

En este artículo se han ofrecido algunas ideas prácticas que, añadidas a las que ya se aplican, reducirían el nivel de ruido en los procesos de revisión por pares, como: usar directrices, emplear una terminología adecuada y consensuada, o adoptar el método de estructuración de juicios complejos. Es cierto que las estrategias de reducción del ruido (higiene de las decisiones) pueden ser costosas en ocasiones (lo primero es vencer el fenómeno de la resistencia al cambio). Sin embargo, la mayoría de las veces sus costes son una mera excusa (Kahneman et al., 2021). En suma, hay que comparar los beneficios de la reducción del ruido con sus costes, y decidir si compensa. En el caso de la evaluación de artículos científicos parece que así puede ser, ya que las medidas que se han propuesto son asequibles y el beneficio esperado es grande porque se conseguiría mejorar el proceso de toma de decisiones, hacerlo más justo y posiblemente lograr una ciencia mejor. En conclusión, el primer paso para resolver el nivel de ruido en los procesos de revisión por pares es reconocer que este problema existe y que es posible medirlo. Este artículo ha sido una primera aproximación a ello.

En cuanto a las limitaciones de nuestro estudio, no hemos tratado la cuestión interesante de cómo detectar y eliminar los sesgos en la revisión por pares, porque requiere un tratamiento diferente al que hemos expuesto aquí (no hay que confundir los conceptos de sesgo y ruido). Por otra parte, si todos los revisores/as hubieran podido evaluar más de un manuscrito, podríamos haber analizado otros aspectos del ruido, como el "ruido de nivel" que se refiere, en nuestro caso, a la variación entre los evaluadores en su disposición a emitir valoraciones más o menos severas sobre la calidad de los manuscritos considerados (es sabido que no todos los *referees* son igual de duros). O el "ruido de patrón", un aspecto más complejo del análisis del ruido que hace referencia a que los evaluadores no son igualmente de severos en todos sus juicios de los manuscritos que revisan: son más duros que su media personal con algunos manuscritos y más indulgentes con otros, lo que refleja un patrón complejo en las actitudes de los jueces hacia casos particulares (el término estadístico más propio para el ruido de patrón es interacción evaluador × caso). Ambos tipos de ruido, de nivel y de patrón, proporcionarían la medición del ruido total del sistema y permitiría realizar una auditoría del ruido completa.

Esto es, se podría diseñar una investigación que tuviera como objetivo estimar el nivel de ruido total en los procesos de evaluación por pares, elaborando el material necesario que debería incluir mediciones cuantitativas de las variables consideradas (lo que daría respuesta a algunas de las preguntas que han planteado autores muy relevantes en el ámbito de la revisión por pares como, por ejemplo Weller, 2001). En este estudio que hemos llevado a cabo, y que podemos entender como piloto, optamos por evaluar un manuscrito previamente aceptado para su publicación, antes de que esta decisión estuviese contaminada por los resultados de la auditoría. En el futuro se podría extender el análisis para observar el ruido en los envíos que han sido rechazados por un editor/a (o un equipo editorial). Así podríamos comprobar no solo el ruido de las personas que revisan, sino también el de los editores de las revistas. Si en lugar de elegir los dos revisores/as que ha tenido un manuscrito rechazado, hubiera tenido otra pareja de revisores/as ¿la decisión editorial final hubiera sido la misma? Incluso se podría evaluar el nivel de ruido en el envío de manuscritos que son rechazados directamente por el editor/a y que no pasan a la revisión por pares porque el editor considera que no tienen la calidad científica adecuada ¿qué probabilidad tendría un manuscrito de haber recibido una valoración positiva por parte de los revisores/as de esa revista? Finalmente, no se

dispone de un nivel de referencia o valor de corte que se pueda fijar como el límite más allá del cual se debería considerar inaceptable el ruido en un proceso de revisión por pares. Seguramente sería tentador importar valores de corte derivados de otros ámbitos (Belur et al., 2021; LeBreton y Senter, 2008; Voskuijl y Van Sliedregt, 2002). Pero el establecimiento de esos valores de corte requiere de una adaptación al contexto donde se aplican y, quizás, también de un análisis de cuáles son los niveles de acuerdo o fiabilidad entre evaluadores en las revistas científicas (Benda y Engels, 2011; Bornmann, 2011).

Author Contributions

Ambos autores han colaborado de manera similar en todo el proceso de elaboración de este artículo.

References

Álvarez, S.M.; Maheut, J. (2022). Protocol: Systematic literature review of the application of the multicriteria decision analysis methodology in the evaluation of urban freight logistics initiatives. *WPOM-Working Papers on Operations Management, 13*(2), 86-107. https://doi.org/10.4995/wpom.16780

Ariely, D. (2008). *Las trampas del deseo. Cómo controlar los impulsos irracionales que nos llevan al error.* Ed. Ariel.

Bedeian, A.G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning & Education, 3*(2), 198-216. https://doi.org/10.5465/amle.2004.13500489

Belur, J.; Tompson, L.; Thornton, A.; Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research, 50*(2), 837-865. https://doi.org/10.1177/0049124118799372

Benda, W.G.G.; Engels, T.C.E. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting, 27(*1), 166-182. https://doi.org/10.1016/j.ijforecast.2010.03.003

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1)*,* 197-245. https://doi.org/10.1002/aris.2011.1440450112

Ernst, E., Saradeth, T., & Resch, K. L. (1993). Drawbacks of peer review. *Nature, 363*(6427), 296. https://doi.org/10.1038/363296a0

Fiske, D.W.; Fogg, L. (1990). But the reviewers are making different criticisms of my paper: Diversity and uniqueness in reviewer comments. *American Psychologist, 45*(5), 591-598.

Hirst, A.; Altman, D.G. (2012). Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS ONE, 7*(4), e35621. https://doi.org/10.1371/journal.pone.0035621

LeBreton, J.M.; Senter, J.L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852. http://orm.sagepub.com/cgi/content/abstract/11/4/815

Kahneman, D. (2012). *Pensar rápido, pensar despacio*. Ed. Debate.

Kahneman D.; Rosenfield A.M.; Gandhi L.; Blaser T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review, 94*(10), 38-46.

Kahneman, D.; Sibony, O.; Sunstein, C.R. (2021). *Ruido. Un fallo en el juicio humano*. Ed. Debate.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. Retrieved from https://repository.upenn.edu/asc_papers/43

Marin-Garcia, J.A.; Santandreu-Mascarell, C. (2015). What do we know about rubrics used in higher education? *Intangible Capital, 11*(1), 118-145. https://doi.org/10.3926/ic.538

Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269, https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

Rezaei, A.R.; Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Theoharakis, Vasilis & Voss, Chris & Hadjinicola, George & Soteriou, Andreas. (2007). Insights into Factors Affecting Production and Operations Management (POM) Journal Evaluation. Journal of Operations Management. 25. 932–955. 10.1016/j.jom.2006.09.002.

Voskuijl, O.F.; Van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*(1), 52-62. https://doi.org/10.1027//1015-5759.18.1.52

Weller, A.C. (2001). *Editorial peer review: its strengths and weaknesses*. Ed. American Society for Information Science and Technology.

Appendix 1

Informed Consent. Request for Participation.

We have reached a decision regarding your submission to WPOM-Working Papers on Operations Management. Our decision is: accept the submission.

In addition, as editor, I am collaborating in a study on noise and bias in the evaluation of scientific articles. If the authors have no objection (and no special urgency for publication of the article), I would like to propose their article for the pilot study we are conducting.

The editorial decision to accept has been made and it will be published in Volume 2 of 2022 in July this year. What we ask you to do is to leave the accepted document hidden until that date (not to make it visible at the time it is typeset, as we usually do). This way it does not interfere with the research pilot.

The pilot consists of sending your paper to several reviewers of the journal and comparing the results of their evaluations (always in a double-blind process and without the reviewers knowing the editorial decision taken).

As an advantage for the authors, their work will be more widely disseminated because it will be seen by many reviewers and, in addition, they will receive comments from multiple people, which may help them to improve future research on the topic.

Moreover, their paper will be appropriately cited in the publication resulting from the pilot study.

Of course, neither your refusal to participate in the pilot study, nor the feedback you may receive from the reviewers, can alter the editorial decision which has already been taken under the normal review conditions of the journal.

I would be grateful if you could let me know whether or not you wish to participate (either way) in order to proceed with the pilot study or to seek a substitute manuscript.

Anexo 1

Consentimiento informado. Solicitud de participación.Hemos llegado a una decisión con respecto a su envío a WPOM-*Working Papers on Operations Management.* Nuestra decisión es: aceptar el envío.

Por otra parte, como editor, estoy colaborando en una investigación sobre ruido y sesgo en la evaluación de artículos científicos. Si los autores no tienen inconveniente (ni una urgencia especial por la visibilidad del trabajo), me gustaría proponer su trabajo para el piloto que estamos llevando a cabo.

La decisión editorial está tomada (es aceptado) y se publicará en el volumen 2 de 2022 en julio de este año. Lo que les solicitamos es dejar oculto el documento aceptado hasta esa fecha (no hacerlo visible en el momento que esté maquetado, como solemos hacer habitualmente). Así no interfiere en el piloto de la investigación.

El piloto consiste en mandar su trabajo a varios revisores de la revista y comparar los resultados de sus evaluaciones (siempre en proceso doble ciego y sin que los revisores sepan la decisión editorial tomada).

Como ventaja para los autores, su trabajo tendrá mayor difusión porque va a ser visto por un número elevado de revisores y, además, recibirán comentarios de múltiples personas, que pueden ayudarles a mejorar investigaciones futuras sobre el tema.

Por otra parte, su artículo será convenientemente citado en la publicación que se derivará del piloto.

Por supuesto, ni la negativa a participar en el piloto por su parte, ni las opiniones que puedan recibirse de los revisores pueden alterar la decisión editorial que ya ha sido tomada en las condiciones normales de revisión de la revista.

Les agradecería que me hicieran llegar su deseo o no de participar (cualquiera de las dos cosas) para proceder con el piloto o buscar un manuscrito sustituto.