Scientometrics (2024) 129:1231–1253
https://doi.org/10.1007/s11192-023-04915-y

How do referees integrate evaluation criteria into their overall judgment? Evidence from grant peer review

Sven E. Hug[1]

Received: 10 December 2022 / Accepted: 8 December 2023 / Published online: 25 January 2024

Abstract
Little is known whether peer reviewers use the same evaluation criteria and how they integrate the criteria into their overall judgment. This study therefore proposed two assessment styles based on theoretical perspectives and normative positions. According to the case-by-case style, referees use many and different criteria, weight criteria on a case-by-case basis, and integrate criteria in a complex, non-mechanical way into their overall judgment. According to the uniform style, referees use a small fraction of the available criteria, apply the same criteria, weight the criteria in the same way, and integrate the criteria based on simple rules (i.e., fast-and-frugal heuristics). These two styles were examined using a unique dataset from a career funding scheme that contained a comparatively large number of evaluation criteria. A heuristic (fast-and-frugal trees) and a complex procedure (logistic regression) were employed to describe how referees integrate the criteria into their overall judgment. The logistic regression predicted the referees' overall assessment with high accuracy and slightly more accurately than the fast-and-frugal trees. Overall, the results of this study support the uniform style but also indicate that the uniform style needs to be revised as follows: referees use many criteria and integrate the criteria using complex rules. However, and most importantly, the revised style could describe most—but not all—of the referees' judgments. Future studies should therefore examine how referees' judgments can be characterized in those cases where the uniform style failed. Moreover, the evaluation process of referees should be studied in more empirical and theoretical detail.

Keywords Academic peer review · Peer review theory · Fast and frugal heuristics · Commensuration bias · Evaluation criteria · Grant funding

✉ Sven E. Hug
sven.hug@uzh.ch

[1] Department of Psychology, University of Zurich, Binzmühlestrasse 14/13, 8050 Zurich, Switzerland

Introduction

In academia, the attribution of value and worth as well as the allocation of reward usually involves peer review processes (Hamann & Beljean, 2017; Zuckerman & Merton, 1971). Peer review is thus seen as central to the practice of research (Baldwin, 2020; Johnson & Hermanowicz, 2017) and pervades modern scholarship.[1] For example, peer review is used in grant funding, scholarly communication (journal articles, books, conference contributions), preregistration of studies, hiring and promotion processes, institutional evaluations, or in the conferral of awards and honors. Given the ubiquity and centrality of the process, it is not surprising that a large literature on peer review has been published (Batagelj et al., 2017; Grimaldo et al., 2018). But despite more than 50 years of research, our theoretical understanding of peer review remains very limited (Bornmann, 2008; Chubin & Hackett, 1990; Gläser & Laudel, 2005; Hirschauer, 2004; Hug, 2022; Reinhart & Schendzielorz, 2021). For instance, we have made little progress with respect to the following broad questions: Why and how has peer review evolved? What are the causal mechanisms that produce peer review phenomena, most notably the many supposed or confirmed biases? How are peer review and the broader contexts in which it is embedded interrelated (organizations, scientific communities, science, society, economy, government)? Why and how does peer review work, and what are its effects? Which interventions and innovations improved peer review, how did they work, and for whom?[2] Due to this theoretical deficit, Hug (2022) encouraged peer review researchers to be more theoretically engaged and to theorize the many aspects of peer review.

The current study focuses on the broad question of how peer review works and, more specifically, on the evaluation process employed by reviewers/referees.[3] With respect to this process, little is known about whether referees use the same evaluation criteria (Arvan et al., 2022; Hug & Ochsner, 2022). Moreover, recent research has raised the question of how referees convert scores along heterogenous evaluation criteria into an overall scale of value (Erosheva et al., 2020; Heesen, 2019; Lee, 2015), or more generally, how referees integrate evaluation criteria into their overall judgment. In this study, I therefore focused on the following research questions: Do referees use the same criteria in their evaluation process? How do referees integrate the criteria into their overall judgment? Specifically, I linked these two questions with theoretical perspectives and normative positions, that is, the three-phase model of editorial judgment, the ideal of impartiality, commensuration bias, and fast-and-frugal heuristics. From these perspectives and positions, I derived antithetical propositions, grouped them into two judgment styles, and tested the two styles empirically using data from a career funding scheme.

[1] The central role of peer review has been re-emphasized in a widely supported European initiative to reform research assessment practices (European CoARA, 2022; European Commission, 2021).

[2] There are many other questions on which we have made little progress, for example: How can we compare and contrast the many facets of peer review across contexts (disciplines, purposes, regions, etc.)? What are the purposes and roles of peer review, and how do peer review processes (not) achieve them? How can we define and conceptualize peer review, given that it is a highly diverse practice, is institutionalized in various ways, includes many different procedures, serves different purposes, and evolves through time?

[3] I use the terms *reviewer* and *referee* interchangeably. A referee is a researcher who contributes an independent and external evaluation of the work of fellow researchers (e.g., a research proposal, a registered report, a manuscript) to a peer review process.

The three-phase model of editorial judgment

Hirschauer (2005, 2010) introduced the three-phase model of editorial judgment based on an empirical analysis of the process used by individual editors who read and assess a manuscript before the editorial meeting. The three phases are characterized by different social frames of reference: The first phase refers to the editor's research community that shares expectations about specific types of texts or genres. The second phase refers to the editor's spontaneous impression, or judgment, that she/he has developed over the course of reading the manuscript. And the third phase represents the editor's written, rationalized judgment on the manuscript by which she/he aims to persuade the other editors. Although the three-phase model focuses on editors, I suggest that it can be generalized to the evaluation process of reviewers because the three phases are similar to the typical process of external refereeing in which peers read and assess, for example, a grant proposal or a manuscript and then provide written reviews to the editor or funding committee.

Hirschauer made several claims related to the second phase of the model that are relevant to the two research questions of this study. Specifically, Hirschauer (2004, 2015) claimed that there is a consensus on evaluation criteria among peers (i.e., in the first phase) but that this consensus quickly dissolves when criteria are practically applied (i.e., in the second phase). As his other claims related to the second phase (Hirschauer, 2004, 2015, 2019) echo Scarr (1982), I present her original position here. According to Scarr, reviewers' and editors' judgments are similar to other complex judgments, for example, those about wine, floral scents, or perfumes. Scarr argues that these judgments are based on complex weightings of many criteria and that the weights are adjusted for the criteria that apply especially to an individual case and the priorities of the referee. Accordingly, she notes that the weightings of criteria can be only partially specified and that judgments are not the result of a simple sum of component parts nor of a simple list of criteria that were combined mechanically. In line with this, Scarr (1982) refers to a referee's evaluation as a "complex human judgment" and Hirschauer (2010) as a "spontaneous impression" and a "spontaneous expression of taste". Based on Hirschauer and Scarr, the following propositions can be derived with respect to the research questions of this study: referees use different evaluation criteria; they use many criteria in their assessment; and they weight, or integrate, the criteria in a complex, non-mechanical way and on a case-by-case basis. The next two sections present theoretical perspectives and normative positions that stand in opposition to Scarr and Hirschauer.

The ideal of impartiality and commensuration bias

According to Lee et al. (2013), the ideal of impartiality implicitly underlies quantitative research on bias in peer review and requires that, among other things, referees "interpret and apply evaluative criteria in the same way in the assessment of a submission" so that they arrive at identical evaluations (pp. 4–5). In line with the ideal of impartiality, Forscher et al. (2019) and Arvan et al. (2022) presuppose that referees apply the same criteria when assessing the same manuscript or grant proposal. In particular, Forscher et al. (2019) argue that if referees do not agree on criteria and use different criteria, this will result in arbitrary and unreliable judgments. And Arvan et al. (2022) contend that only if referees and potential readers of a paper agree on what constitutes quality, the judgments of referees are useful to readers when deciding how to allocate their reading time among papers. In

contrast to Scarr and Hirschauer, these authors state that referees are supposed to use the same evaluation criteria.

Building on the ideal of impartiality, Lee (2015) proposed commensuration bias as a new type of bias. Lee conceptualizes *commensuration* as the process by which referees convert "a submission's strengths and weaknesses for heterogeneous peer review criteria into a single metric of quality or merit" (p. 1272) and defines *commensuration bias* as referees' "deviation from the impartial weighting of peer review criteria in determinations of a submission's final value" (p. 1273). Lee argues that the deviation from impartial weighting can have three sources, one of which, referee idiosyncrasy, is relevant to the present study.[4] Referee idiosyncrasy means that referees use their own, idiosyncratic weightings and vary weightings across proposals or contexts. Idiosyncrasy is consistent with the way Scarr and Hirschauer characterize the evaluation process employed by referees (see 1.1). According to Lee, however, referees are supposed to integrate evaluation criteria into their overall judgment with the same weightings (i.e., the same weightings as the other referees and the same weightings across proposals) to avoid commensuration bias.

Fast-and-frugal heuristics

Heuristics are generally defined as strategies that enable decision makers to process information in a less effortful manner than one would expect from an optimal decision rule (Shah & Oppenheimer, 2008). More specifically, in the research program on fast-and-frugal heuristics, a heuristic is defined as a procedure for making decisions under uncertainty that ignores a part of the available information in order to make decisions more quickly, frugally, and/or accurately than complex procedures (Gigerenzer et al., 2022).[5] Four concepts in this definition need further clarification: uncertainty, frugality, speed, and complex procedures. *Uncertainty* refers to "situations in which perfect foresight of all future events, their consequences, and probabilities is impossible" and in which "the optimal decision cannot be determined" (Gigerenzer et al., 2022, pp. 172 and 174). The focus on uncertainty makes fast-and-frugal heuristics particularly useful for studying decision-making in grant funding and peer review because the outcomes of proposed research are unpredictable and judgments on research quality are associated with a high degree of uncertainty (Bornmann, 2015). *Frugality* refers to the number of cues a heuristic uses. The less cues a heuristic uses, the more frugal it is (Gigerenzer et al., 1999). In the present study, evaluation criteria represent cues. *Speed* refers to the number of operations that need to be performed to make a decision. The less computation a heuristic needs, the faster it is (Gigerenzer et al., 1999), and more frugal heuristics enable faster decisions (Wegwarth et al., 2009). *Complex procedures* assess and use all available information (i.e., they are lavish and not frugal) and integrate all information in an optimal and computationally expensive way (i.e., they are

[4] In addition to referee idiosyncrasy, Lee (2015) mentions two other sources for commensuration bias. First, referees weight criteria according to social characteristics of the applicants or authors. Second, referees undervalue criteria that promote truth and innovation in science (e.g., methodological soundness, novelty).

[5] The research program on fast and frugal heuristics should not be confused with the *heuristics and biases approach* (Gilovich et al., 2002; Tversky & Kahneman, 1974), which, in contrast to the fast-and-frugal program, favors a skeptical attitude towards human judgment (Kahneman & Klein, 2009) and has identified heuristics as a source of biases and errors (Gilovich et al., 2002; Tversky & Kahneman, 1974).

slow and not fast) (Gigerenzer & Goldstein, 1996). The literature on fast-and-frugal heuristics often mentions regression models, Bayesian models, and artificial neural networks as examples of complex procedures.

To illustrate how fast-and-frugal heuristics work, I briefly describe one such heuristic. The take-the-best heuristic (Gigerenzer & Goldstein, 1996) decides between two alternatives by using the cue with the highest cue validity. Martignon and Hoffrage (1999) defined cue validity as the accuracy with which a cue predicts a judgment. If no discrimination can be made, the next best cue is used, and so on. In numerous studies, the performance of this and other fast-and-frugal heuristics has been compared to the performance of complex procedures using empirical data or simulations. For example, Czerlinski et al. (1999) compared the performance of the take-the-best heuristic to the performance of multiple linear regression across 20 real-world environments using cross-validation (training, testing). They found that the heuristic was slightly more accurate in the testing set than the regression (71 vs 68% correct results), while using considerably fewer cues (2.4 vs 7.7 cues on average).[6] One of the key findings of this research is the less-is-more effect: less information and computation can lead to more accurate judgments than more information and computation (Gigerenzer & Goldstein, 1996; Gigerenzer et al., 2022). The findings from the fast-and-frugal programm have the following implications for the research questions of this study: referees use a small number of the available criteria in their assessment, and they integrate the criteria into their overall judgment based on simple rules that are computationally inexpensive.

The current study

The antithetical propositions presented in the previous sections can be grouped into a *case-by-case* and a *uniform* style of assessment/judgment. According to the case-by-case style, referees use many and different evaluation criteria, weight criteria on a case-by-case basis, and integrate criteria in a complex, non-mechanical way into their overall judgment. "Case-by-case basis" means that referees use a distinct weighting of criteria for every proposal. According to the uniform style, referees use a small fraction of the available criteria, apply the same criteria, weight the criteria in the same way (i.e., same weightings as the other referees and the same weightings across proposals), and integrate the criteria based on simple rules (i.e., fast-and-frugal heuristics).[7] These two styles have not been examined in research on peer review, but some studies on grant funding contain evidence for the uniform style. Specifically, studies that regressed the referees' overall assessment scores on the criteria scores found that most or all criteria are positively related to the overall scores (Eblen et al., 2016; Erosheva et al., 2020; Lindner et al., 2016; Rockey, 2011; Würth

[6] The results, however, were the other way around in the training set: the regression was slightly more accurate (77% correct) than the heuristic (75% correct). This very pattern is often observed in studies on fast-and-frugal heuristics: complex procedures overfit, while heuristics avoid overfitting due to their simplicity and thus outperform complex procedures in the testing set (Artinger et al., 2022; Gigerenzer et al., 1999, 2011). The robustness and generalizability of fast-and-frugal heuristics is thus often higher than that of complex procedures.

[7] The case-by-case style and the uniform style resemble Meehl's distinction between clinical and statistical judgment (Meehl, 1954). Note that the case-by-case and uniform style are both conceived as human judgments, while Meehl only sees clinical judgment as a judgment that is formed "in the head" of humans (Grove & Meehl, 1996, p. 293). From Meehl's perspective, both styles would be considered clinical because they are formed by humans (Grove, 2005; Meehl, 1954).

et al., 2017). Moreover, a subset of these studies reported that the criteria scores explain the variability of the overall scores to a large extent (Eblen et al., 2016; Erosheva et al., 2020; Lindner et al., 2016). The results of these studies suggest that referees use the same criteria, weight them in the same way, and integrate them mechanically using a complex procedure (i.e., linear regression). However, these studies have not investigated whether referees make fast and frugal judgments (i.e., whether they use a small fraction of the available criteria and integrate the criteria based on simple rules). And the studies analyzed data from funding schemes that employed a low number of evaluation criteria (i.e., three to five), which might have concealed that referees use many and different evaluation criteria, as suggested by the case-by-case judgment style.

The current study therefore examined the case-by-case and uniform style of judgment using a unique dataset from a career funding scheme that contained a comparatively large number of evaluation criteria. A heuristic (fast-and-frugal trees) and a complex procedure (logistic regression) were employed to describe how referees integrate the criteria into their overall judgment. Based on the case-by-case judgment style, I expected that a fast-and-frugal tree or regression equation could not be identified—or if they could be identified, I expected that the criteria scores would predict the overall judgment with very low accuracy—because referees use many different criteria and integrate them in a non-mechanical way and on a case-by-case basis. Based on the uniform style, I expected that a fast-and-frugal tree and regression equation could be specified and that the criteria would predict the overall judgment with high accuracy because referees use the same criteria and integrate them using the same mechanical rules and the same weightings. If a tree and regression could be specified, I expected, based on the fast-and-frugal literature, that the tree would use a small number of the evaluation criteria and deliver more accurate predictions of the referees' overall judgment than the regression that includes all evaluation criteria. In line with the fast-and-frugal paradigm, I assessed the performance of the tree and regression in a cross-validation design in which the data was split into a training set and testing set.

Many fast-and-frugal heuristics are available, and this study employed one particular heuristic, fast-and-frugal-trees, for the following reasons.[8] Unlike the class of one-clever-cue heuristics (Gigerenzer et al., 2022), fast-and-frugal trees are flexible in terms of the number of cues included because they can be created with just one cue or expanded to include more cues (Martignon et al., 2003). They are therefore suited to examine the research questions of this study. Moreover, fast-and-frugal trees are lexicographic and thus non-compensatory, that is, the cues are used in order of their cue validity and later cues thus cannot reverse the decision made by an earlier cue (Gigerenzer et al., 1999). For example, if a grant proposal is considered unoriginal (first cue) and therefore rejected, its methodological strength (second cue) or feasibility (third cue) cannot compensate for this because later cues remain unconsidered in a fast-and-frugal tree. In contrast, complex procedures such as regression models can compensate unoriginality by methodological rigor or feasibility because they integrate all cues. Fast-and-frugal trees thus use a markedly different

[8] Naturally, there are also many different complex procedures available. However, this study employed logistic regression because regression models are a popular method and benchmark in studying human judgment, including peer review. For example, regression models have been used for decades to directly study and describe human judgment and behavior in a variety of contexts (Beckstead, 2007; Gigerenzer et al., 2011), including peer review (e.g., Lindner et al., 2016; Porter & Rossini, 1985; Reinhart, 2009). Moreover, regression models are often used as benchmark in fast-and-frugal studies (Gigerenzer et al., 2011) and sometimes serve as benchmark in peer review studies (e.g., Devyatkin et al., 2018; Kang et al., 2018).

type of integration logic than complex procedures, which might be useful for understanding peer review judgments in a novel way. Lastly, fast-and-frugal trees have been studied in a variety of contexts, including medical, legal, financial, and managerial decision-making (Phillips et al., 2017), but not in the domain of peer review.[9]

Methods

Data

The data consisted of 474 rating forms on 237 proposals that were submitted to a funding scheme for doctoral students and postdocs in two consecutive years. The purpose of the scheme is to support outstanding early career researchers from all disciplines who need funding to start or complete their research project. The unnamed funding organization relies on a small and stable pool of referees representing all disciplines to assess the applications; the referees are thus well acquainted with the evaluation procedure and the use of the rating form. From this pool, 31 referees rated the 237 proposals. The review load was unevenly distributed. Six referees each contributed more than 5% of the total reviews and accounted for 45% of all reviews; thirteen referees each contributed between 2 and 5% of all reviews, accounting for 41% of all reviews; twelve referees each contributed less than 2% of all reviews, accounting for 14% of all reviews. Every application was assessed by two referees (a first referee and a second referee). The referees rated an application on 13 criteria and provided an overall assessment (see Table 1). The majority of the evaluation criteria were scored on a five-point scale (1 poor, 2 average, 3 good, 4 very good, 5 excellent). However, the criteria "career potential" (i.e., a combination of academic potential, resilience, and long-term academic interest) and "overall assessment" included an additional scoring option (6 outstanding), while the criteria "applicant belongs to the top-5% in the field" and "proposal is highly innovative" had binary response options (yes/no). A unique feature of the present dataset is the high number of evaluation criteria, which makes it particularly suited to test the research questions. In comparison, NIH referees score proposals on five criteria (Erosheva et al., 2020), while proposals in the EU's Marie Skłodowska-Curie Actions are scored on three criteria (Pina et al., 2021), and project applications at the Swiss National Science Foundation are rated on three criteria (Würth et al., 2017).

In the fast-and-frugal paradigm, decision rules are assessed using cross-validation to reduce overfitting and to increase generalizability (Czerlinski et al., 1999; Gigerenzer & Gaissmaier, 2011; Gigerenzer et al., 1999; Wang et al., 2022). In line with this approach, the data was split into a training set consisting of the ratings from the first referees ($n = 237$) and a testing set consisting of the ratings from the second referees ($n = 237$). The descriptive statistics of the training and testing set are shown in Table 1. As fast-and-frugal trees are designed for binary classification and decision tasks (Martignon et al., 2003; Phillips et al., 2017) and as no suitable binary dependent variable was included in the dataset, the "overall assessment" was binarized (not outstanding = scores 1 to 4; outstanding = scores 5 or 6). This binarization can be supported by two arguments. First, the scores 5 and 6 are

[9] While the present study is the first that models peer review judgments using fast-and-frugal heuristics, fast-and-frugal heuristics have already been employed in the domain of research evaluation to describe and prescribe the use of bibliometric indicators (e.g., Bornmann & Hug, 2020; Bornmann & Marewski, 2019; Bornmann et al., 2022; de Abreu Batista Júnior et al., 2021).

referred to as "excellent" and "outstanding", respectively, when used as separate scores in the rating forms. However, they also have a common, superordinate meaning in the evaluation procedure and signify "absolutely necessary to be funded". Second, the funding rate of the funding scheme was 51.8% in the analyzed two years, which is almost identical to the binarized "overall assessment" of the those acting as first referee (50.6% outstanding) and somewhat higher than the binarized "overall assessment" of the those acting as second referee (45.6% outstanding). The difference of 5% between the first and second referees might be a coincidence, or it might be due to the selection procedure of the funding organization, which appoints the referee who has the most subject-matter expertise on a proposal as first referee. However, the latter explanation is unlikely as the few empirical studies on this topic suggest the opposite: the closer a referee is to a proposal, the harsher the evaluations (for a discussion, see Gallo et al., 2016).

Statistical analysis

Fast-and-frugal trees consist of cues sequentially ordered by cue validity (i.e., from high to low) and binary decisions based on these cues (Martignon et al., 2003, 2008). A tree can be represented graphically as a decision tree or described verbally by a series of if–then-else statements. For example, if a grant proposal is considered unoriginal then reject it; otherwise, examine whether it is methodologically sound. If it is not sound, reject it; otherwise, examine whether it is feasible. If it is not feasible, reject it; otherwise, fund it. In the current study, fast-and-frugal trees were constructed and assessed in R 4.1.2 (R Core Team, 2021) using the FFTrees package 1.6.6 (Phillips et al., 2017). The dataset generated by the first referees (training set) was used to fit fast-and-frugal trees with one to six cues (six is the maximum number of cues the FFTrees package can fit). The 13 evaluation criteria served as cues and the binarized "overall assessment" as criterion variable. In line with the requirement to ignore conditional dependencies between cues when building fast-and-frugal trees (Martignon et al., 2003), the ifan algorithm was used to create the trees. From the trees created by the ifan algorithm, those with the highest accuracy were selected. For cross-validation, the performance of the resulting six fast-and-frugal trees was tested using the dataset generated by the second referees (testing set). In addition to the fast-and-frugal trees, a logistic regression was computed. Specifically, the binarized "overall assessment" was regressed on the 13 evaluation criteria in the dataset generated by the first referees (training set) using the glm function and the binomial family implemented in R 4.1.2 (R Core Team, 2021). The cutoff probability to classify the overall judgment as "outstanding" and "not outstanding" was determined using the argument "misclasserror" in the optimalCutoff function from the InformationValue package 1.2.3 (Prabhakaran, 2016), which minimizes the misclassification rate. Collinearity was assessed exploiting the vif function from the car package 3.1.0 (Fox & Weisberg, 2019). For cross-validation, the resulting regression model was then applied to the dataset generated by the second referees (testing set). In addition to regressing all 13 criteria on the overall judgment, two regressions with the most important criteria only were computed to assess how "frugal" regressions perform. The most important criteria were selected in two ways. On the one hand, the same criteria were used in the regression as in the best performing fast-and-frugal tree (i.e., the criteria with the highest cue validities). On the other hand, the "significant" criteria from the full regression were included using a traditional frequentist statistics cutoff of $p = 0.05$ (i.e., the criteria with the highest log-odds).

Table 1 Descriptive statistics of the proposal ratings (training set, $n = 237$; testing set, $n = 237$)

Evaluation criteria (cues)	Training set				Testing set			
	M	SD	Mdn	min, max	M	SD	Mdn	min, max
Applicant								
Education	4.4	0.7	4	2, 5	4.3	0.8	4	1, 5
Track record	4.1	0.9	4	1, 5	4.1	0.9	4	1, 5
Career plan	4.1	0.9	4	2, 5	3.9	1.0	4	1, 5
Career potential	4.5	1.1	5	1, 6	4.3	1.0	4	2, 6
Top 5% in the field (% yes)	14%				13%			
Project								
Clarity of project goal	4.5	0.7	5	2, 5	4.5	0.7	5	2, 5
Clarity of research plan	4.4	0.8	5	1, 5	4.3	0.9	5	2, 5
Own research question	4.4	0.8	5	1, 5	4.2	1.0	5	1, 5
Methodological approach	4.4	0.8	5	1, 5	4.3	1.0	5	1, 5
Feasibility	4.2	0.9	4	1, 5	4.1	1.0	4	1, 5
Highly innovative proposal (% yes)	19%				11%			
Environment								
Cooperation, network	4.4	0.8	5	2, 5	4.5	0.7	5	2, 5
Recommendation letter	4.4	0.7	5	1, 5	4.5	0.7	5	2, 5
Overall assessment	4.4	1.0	5	1, 6	4.2	1.2	4	1, 6

The performance of the trees and regression models was evaluated based on five metrics. *Absolute frugality* (#Frug) is the average number of cues a decision rule integrates to reach a decision. For example, a heuristic that consists of two cues and uses one cue to reach a decision in 40% of all decisions and two cues in 60% of all decisions has an absolute frugality of 1.6. *Relative frugality* (%Frug) represents the share of cues that are ignored and equals 1 minus the absolute frugality divided by all available cues. For example, a heuristic ignores 88% of the cues if 13 cues are available and the heuristic uses 1.6 cues on average.[10] In addition to the frugality metrics, three metrics based on the confusion matrix were used. *Accuracy* (Acc) is the sum of true positives and true negatives divided by the number of all cases or decisions. It is the most widely used metric in research on fast-and-frugal heuristics. *Sensitivity* (Sens) is the number of true positives divided by the sum of true positives and false negatives. *Specificity* (Spec) is the number of true negatives divided by the sum of true negatives and false positives.

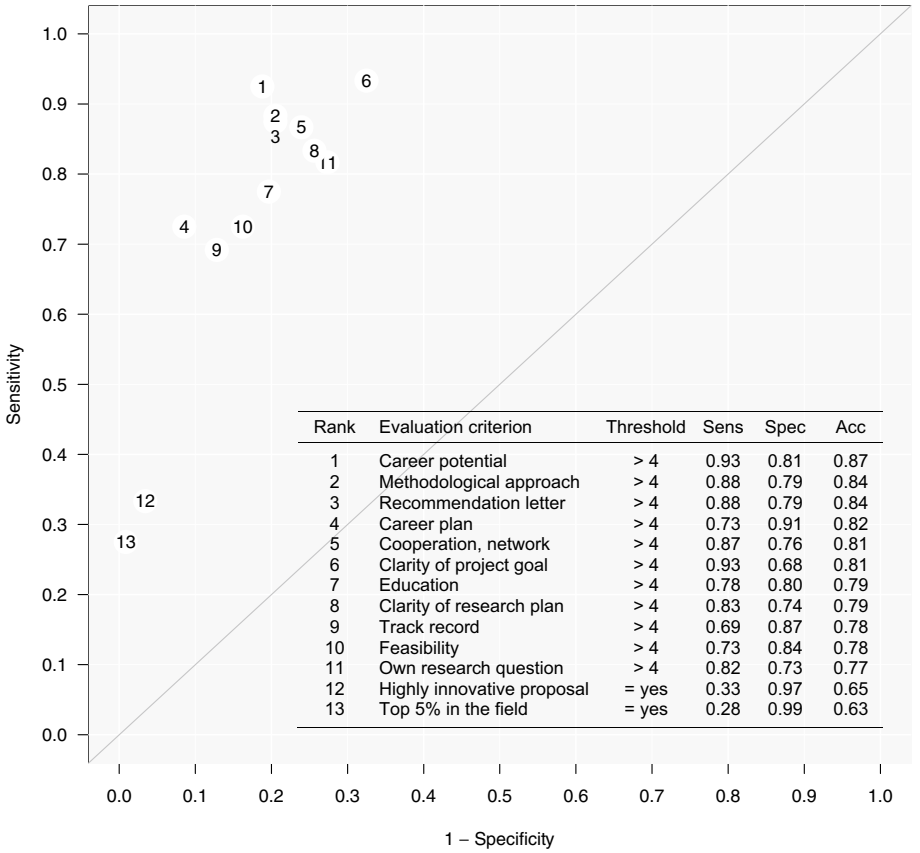
[10] I use the same two metrics as Phillips et al. (2017) (i.e., mean cues used; percent cues ignored) but name them differently (i.e., absolute frugality; relative frugality). Moreover, I interpret the metric "mean cues used" differently. While Phillips and colleagues refer to "mean cues used" as a speed measure, I interpret it as a measure of frugality (i.e., absolute frugality), which is line with Czerlinski et al. (1999) who defined frugality as the "average number of cues looked up" (p. 103).

Results

The cue validities of all thirteen evaluation criteria, which were calculated from the ratings in the training set ($n = 237$), are shown in Fig. 1. Cue validity is defined as the accuracy with which a cue predicts a judgment; a cue is informative and valid for making a judgment if its accuracy is > 0.5 (Martignon & Hoffrage, 1999). According to Fig. 1, the accuracy of all cues is > 0.5, which indicates that all cues are valid and, more broadly, that all evaluation criteria are related to the overall judgment. The "career potential" of the applicant has the highest cue validity (0.87), while the criterion "applicant belongs to the top-5% in the field" has the lowest cue validity (0.63). Figure 2 shows fast-and-frugal-trees describing the referees' overall judgments in the training set ($n = 237$) with one, two, and three cues (i.e., evaluation criteria). According to these trees, the "career potential" of the applicant is the most important evaluation criterion for referees, which is consistent with the goal of the funding scheme to promote outstanding early career researchers. The two other evaluation criteria, "methodological approach" and "letter of recommendation", are placed at lower levels of the trees due to their lower cue validity.

The tree with two cues (Fig. 2b) indicates that a referee assesses a proposal as outstanding only if the applicant's career potential is excellent or outstanding (scores 5 and 6, respectively) and the methodological approach is excellent (score 5). The tree thus represents a conjunctive decision rule and should minimize false positives (Einhorn, 1970; Gigerenzer et al., 2022). In fact, the tree demonstrates high specificity and lower sensitivity (Table 3) and generated a total of 14 false positives and 49 false negatives ($N = 474$). In contrast, the tree with three cues (Fig. 2c) represents a zigzag pattern and should thus minimize both false positives and false negatives (Gigerenzer et al., 2022; Martignon et al., 2003). As Table 3 demonstrates, sensitivity and specificity are more balanced in this tree than in the two-cue tree. More specifically, the tree with three cues produced a total of 29 false positives and 37 false negatives ($N = 474$). The tree indicates that a referee does not evaluate an application unfavorably if the methodological approach is not sufficient (scores 1–4), but she/he considers further information, the letter of recommendation, to make the final judgment (Fig. 2c). As research on fast-and-frugal heuristics prefers results from the testing set to results from the training set due to greater generalizability (Gigerenzer et al., 1999; Wang et al., 2022), the tree with two cues in the testing set and an accuracy of 0.85 can be considered the one that best describes the judgment process of referees although it is only marginally more accurate than the other trees in the testing set (Table 3). This tree uses on average 1.4 cues in the testing set to predict the referees' overall judgment and thus ignores 89% of the available cues (training set: #Frug = 1.6, %Frug = 0.88). A verbal description of the trees consisting of four, five, and six cues is provided in the Appendix (Table 4).

Table 2 shows the results of the logistic regression, in which the referees' overall judgments in the training set ($n = 237$) were regressed on all thirteen evaluation criteria (AIC = 123.95; X^2 (13, $N = 237$) = 232.56, $p < 0.001$; pseudo-R^2: McKelvey–Zavoina = 0.90, Veall–Zimmermann = 0.85, McFadden = 0.71). According to the estimates in Table 2, all evaluation criteria are positively related to the overall judgment, suggesting that referees use a broad range of criteria to evaluate an application and integrate the criteria in a linear, additive way. However, the confidence intervals of the estimates are wide, likely due to collinearity. Rules of thumb for identifying collinearity suggest that a VIF < 4, < 5, or even < 10 indicates low collinearity. The VIFs reported in Table 2 could thus be considered unproblematic. Johnston et al. (2018), however, argued that even low levels of collinearity



Rank	Evaluation criterion	Threshold	Sens	Spec	Acc
1	Career potential	> 4	0.93	0.81	0.87
2	Methodological approach	> 4	0.88	0.79	0.84
3	Recommendation letter	> 4	0.88	0.79	0.84
4	Career plan	> 4	0.73	0.91	0.82
5	Cooperation, network	> 4	0.87	0.76	0.81
6	Clarity of project goal	> 4	0.93	0.68	0.81
7	Education	> 4	0.78	0.80	0.79
8	Clarity of research plan	> 4	0.83	0.74	0.79
9	Track record	> 4	0.69	0.87	0.78
10	Feasibility	> 4	0.73	0.84	0.78
11	Own research question	> 4	0.82	0.73	0.77
12	Highly innovative proposal	= yes	0.33	0.97	0.65
13	Top 5% in the field	= yes	0.28	0.99	0.63

Fig. 1 Cue validities of the thirteen evaluation criteria in the ROC space (training set, $n = 237$). *Acc* accuracy, *Sens* sensitivity, *Spec* specificity

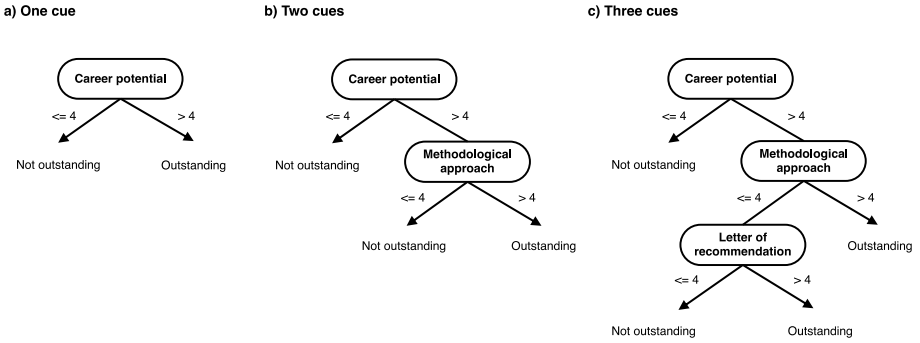


Fig. 2 Fast-and-frugal-trees describing the referees' overall judgment (outstanding, not outstanding) with one, two, and three cues (training set, $n = 237$)

(VIF < 2.5) can lead to unreliable regression coefficients. Moreover, Lindner et al. (2016) regressed the overall assessment of NIH referees on five evaluation criteria and found that the correlated criteria render regression coefficients unreliable. Caution should therefore be exercised when determining the importance of individual evaluation criteria based on the log-odds and p-values reported in Table 2. Although Table 2 does not provide conclusive evidence that referees use all criteria, the estimates, the superior accuracy (Table 3), and the better goodness-of-fit statistics of the regression model including all cues (see statistics of the "frugal" regressions below) suggest that referees do so.

Table 3 indicates that the regression that includes all thirteen cues is slightly more accurate than the fast-and-frugal tree that best describes the referees' judgment process (i.e., the two-cue tree). More specifically, the regression outperforms the tree in terms of accuracy in both training (0.92 vs 0.88) and testing (0.89 vs 0.85). However, in contrast to the tree, the regression is lavish because it uses all available cues (#Frug = 13) and ignores none (%Frug = 0). Hence, much more information and computation produced slightly more accurate judgments than considerably less information and computation.

To assess how "frugal" regressions would perform, two regressions with statistically significant criteria only (referred to as log-odds model) and the criteria from the best performing fast-and-frugal tree (referred to as cue-validity model) were computed. Although it was pointed out above that interpreting the log-odds and p-values is probably not very meaningful, a frugal regression model was defined based on statistical significance because blindly using statistical significance is a typical strategy employed in judgment analysis studies to identify the cues that judges use (Beckstead, 2007). When applying this strategy and using the traditional frequentist statistics cutoff of $p = 0.05$, the two criteria with the highest log-odds in Table 2, "clarity of the project goal" ($p = 0.002$) and "track record" ($p = 0.03$), would be considered the cues that the referees have used. This model, however, is less accurate than the cue-validity model, which included the criteria from the best performing fast-and-frugal tree (i.e., "career potential" and "methodological approach", the cues with the highest cue validities). More specifically, the log-odds model is less accurate than the cue-validity model in both training (0.85 vs 0.89) and testing (0.82 vs 0.85; Table 3). This is also reflected in the goodness-of-fit statistics of the log-odds model (AIC = 161.94; X^2 (2, $N = 237$) = 172.57, $p < 0.001$; pseudo-R^2: McKelvey–Zavoina = 0.78, Veall–Zimmermann = 0.73, McFadden = 0.53), which are less favorable than those of the cue-validity model (AIC = 142.09; X^2 (2, $N = 237$) = 192.42, $p < 0.001$; pseudo-R^2: McKelvey–Zavoina = 0.81, Veall–Zimmermann = 0.77, McFadden = 0.59).

The two frugal regression models provide a baseline against which the performance of the regression including all cues can be assessed. While the frugal models both ignore 85% of the available cues, their accuracy is not considerably lower than that of the model that includes all cues and ignores none (Table 3). In particular, the log-odds and cue-validity model are only slightly less accurate than the model including all cues in both training (0.85 and 0.89 vs 0.92) and testing (0.82 and 0.85 vs 0.89). Using significantly more criteria in a complex procedure thus increased accuracy only slightly.

Since the cue-validity model and the two-cue tree consist of the same criteria (i.e., career potential, methodological approach), the performance of the two decision rules, logistic regression and fast-and-frugal trees, can be directly compared. According to Table 3, the two-cue tree is slightly more frugal than the cue-validity model because it ignores 88% (training) and 89% (testing) of the available cues, whereas the regression ignores 85% of the cues. The tree, however, is much faster (i.e., computationally less demanding) because it performs a few simple comparisons (i.e., "is less than", "is greater than") to make a

Table 2 Logistic regression analysis of the referees' overall judgments (training set, $n = 237$)

Variable	Estimate	*SE*	95% CI		*p*	VIF
			LL	UL		
Applicant						
Education	0.26	0.57	− 0.87	1.40	.66	2.2
Track record	1.42	0.47	0.55	2.40	< .01	2.6
Career plan	0.53	0.43	− 0.31	1.38	.21	2.7
Career potential	0.87	0.49	− 0.07	1.86	.07	3.8
Top 5% in the field	0.73	1.38	− 1.63	4.13	.60	1.4
Project						
Clarity of project goal	1.51	0.68	0.24	2.93	.03	2.4
Clarity of research plan	0.69	0.54	− 0.36	1.79	.21	2.6
Own research question	0.49	0.52	− 0.55	1.52	.35	2.1
Methodological approach	1.01	0.54	− 0.04	2.09	.06	2.9
Feasibility	0.09	0.41	− 0.74	0.89	.83	2.1
Highly innovative proposal	0.50	0.74	− 0.90	2.06	.50	1.3
Environment						
Cooperation, network	0.34	0.53	− 0.69	1.40	.52	2.3
Recommendation letter	0.82	0.46	− 0.08	1.74	.07	2.2
Constant	− 36.15	5.99	− 49.49	− 25.81	< .01	

CI confidence interval, *Estimate* log-odds, *SE* standard error, *LL* lower limit, *UL* upper limit, *VIF* variance-inflation factor

Table 3 Performance of fast-and-frugal trees and logistic regressions in the training set (237 judgments) and testing set (237 judgments)

Decision rule	Training set					Testing set				
	#Frug	%Frug	Acc	Sens	Spec	#Frug	%Frug	Acc	Sens	Spec
Fast-and-frugal trees										
1 cue	1	92%	.87	.93	.81	1	92%	.82	.80	.84
2 cues	1.6	88%	.88	.83	.94	1.4	89%	.85	.74	.95
3 cues	1.7	87%	.90	.90	.91	1.5	88%	.82	.77	.86
4 cues	1.7	87%	.90	.88	.92	1.6	88%	.84	.77	.91
5 cues	1.7	87%	.91	.89	.92	1.6	88%	.84	.77	.90
6 cues	1.8	86%	.91	.89	.92	1.6	88%	.84	.77	.90
Logistic regression										
2 cues (log-odds)	2	85%	.85	.91	.79	2	85%	.82	.91	.75
2 cues (cue validity)	2	85%	.89	.86	.92	2	85%	.85	.74	.94
All cues	13	0%	.92	.95	.89	13	0%	.89	.88	.90

#Frug absolute frugality, *%Frug* relative frugality, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity

decision, while the logistic regression needs to execute multiplications and additions plus a comparison with a cutoff. In terms of accuracy, the tree and the regression are virtually identical in training (0.882 vs 0.890) and testing (0.852 vs 0.848; the respective values are

rounded to two decimals in Table 3). Hence, when the same few evaluation criteria were used, fast-and-frugal trees offered a more parsimonious explanation of the referees' overall judgment than logistic regression.

To assess the robustness of the performance of the fast-and-frugal trees and logistic regressions reported in Table 3, a second training and testing set was created using stratified random sampling. Specifically, the training set contained 50% of the judgments from the first referees and 50% of the judgments from the second referees (total 238 judgments). The testing set contained the remaining 236 judgments (50% from the first referees and 50% from the second referees). The results are provided in the Appendix (Fig. 5). They are consistent with the results in Table 3 and lead to the same conclusions.

To examine whether only a sub-group of the 31 referees used the same decision rule, for example those referees contributing most reviews, and thus account for the results of this study, the share of a referee's total judgments correctly described by the fast-and-frugal tree (2 cues) and the logistic regression (13 cues) were calculated. Figure 3 (training set) and Fig. 4 (testing set) show that the tree and the regression describe a large proportion of each referee's judgments and that the regression's performance is superior to that of the tree. More specifically, the tree correctly describes 80% or more of the judgments of 24 referees in the training set and of 22 referees in the testing set, while the regression correctly describes 80% or more of the judgments of 26 referees in both the training and testing set. Moreover, the regression describes the judgments of eleven referees more accurately than the regression in the training set (referees 2, 3, 5, 6, 7, 8, 11, 12, 16, 19, 25) and of seven referees in the testing set (referees 2, 3, 5, 13, 23, 25, 31). In contrast, the tree describes the judgments of only four referees more accurately than the regression in the training set (referees 1, 10, 14, 26) and of three referees in the testing set (referees 1, 14, 18). In Figs. 3 and 4, the variability of the correctly described judgments increases from left (referees contributing the most reviews) to right (referees contributing the least reviews), which is likely due to the decreasing sample size. Overall, Figs. 3 and 4 suggest that referees mostly use the same decision rule as their fellow referees and the same rule across proposals, providing support for the uniform judgment style. However, Figs. 3 and 4 also clearly demonstrate

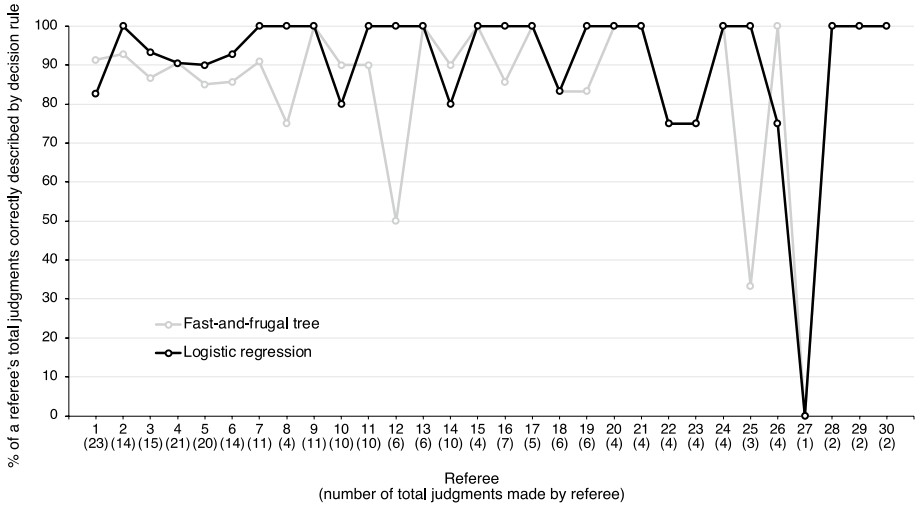


Fig. 3 Training set: Proportion of a referee's total judgments correctly described by the two-cue fast-and-frugal tree and the logistic regression that included all thirteen cues (referees: $n = 30$, judgments: $n = 237$)

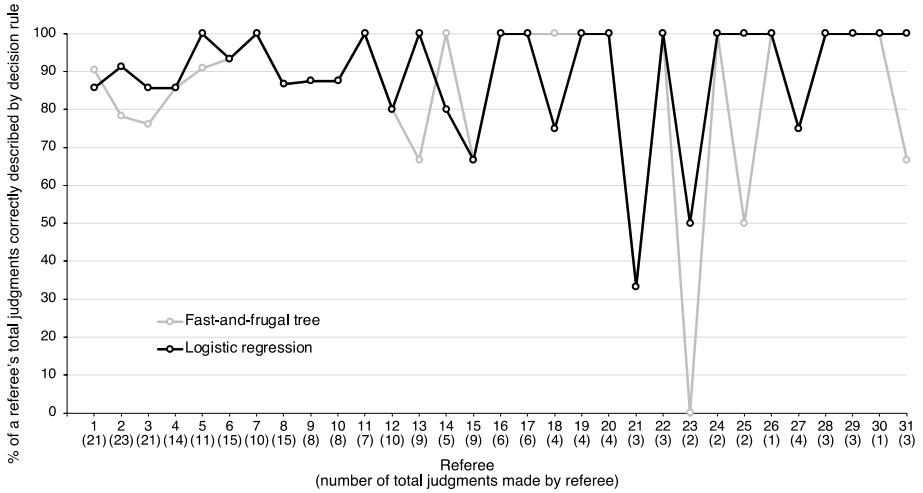


Fig. 4 Testing set: Proportion of a referee's total judgments correctly described by the two-cue fast-and-frugal tree and the logistic regression that included all thirteen cues (referees: $n = 31$, judgments: $n = 237$)

that not all referees apply the same decision rule all the time, which is logically consistent with the high but not perfect accuracy values reported in Table 3.

Discussion and conclusion

The evaluation process that referees employ in academic peer review is undertheorized. With respect to this process, little is known whether referees use the same evaluation criteria and how they integrate the criteria into their overall judgment. This study therefore proposed two assessment styles, the case-by-case and the uniform style, based on theoretical perspectives and normative positions. These styles were tested using data from a career funding scheme for doctoral students and postdocs from all disciplines. The results of this study suggest that referees use many evaluation criteria, apply the same criteria, weight the criteria in the same way (i.e., the same weightings as the other referees and the same weightings across proposals), and integrate the criteria mechanically using a complex rule (i.e., a linear-additive model). This is generally consistent with the uniform judgment style. But, based on the results of this study, the uniform style needs to be revised in two respects. The uniform style proposed that referees use a fraction of the available criteria, while the results suggest that referees use *many criteria*. In addition, the *complex procedure* (logistic regression) was more accurate than the simple rule (fast-and-frugal heuristics) proposed in the uniform style. However, and most importantly, the revised style applies to most—but not all—of the referees' judgments in this study. The fact that the revised uniform style cannot describe all judgments could be due to the following reasons. First, the referees sometimes used a case-by-case judgment style, that is, they weighted the criteria on a case-by-case basis, applied different criteria, or integrated the criteria using other rules. Second, the decision rules employed in this study (logistic regression, fast-and-frugal trees) were unable to reveal the true rules referees apply. Other heuristics or other complex algorithms, such as elastic net regression or support-vector machines, may be better suited to capture

how referees integrate evaluation criteria. Lastly, the referees uniformly used additional criteria not covered by the rating form. For example, referees may have systematically looked up the citation impact of the applicants in bibliometric databases or based their judgment also on bias factors, such as status or gender. Or referees applied universalistic merit criteria that were not included in the rating form.

The results from the logistic regression including all cues are consistent with studies on grant peer review that regressed the referees' overall assessment scores on the criteria scores. Similar to the present study, these studies found that most or all evaluation criteria are positively related to the referees' overall scores (Eblen et al., 2016; Erosheva et al., 2020; Lindner et al., 2016; Rockey, 2011; Würth et al., 2017) and that the criteria scores explain the variability of the overall scores to a large extent but not completely (Eblen et al., 2016; Erosheva et al., 2020; Lindner et al., 2016). These studies, however, used data from funding schemes in which referees scored the proposals on three to five criteria, while the present study analyzed data from a funding instrument that encompassed thirteen criteria. The present study thus extends the results of previous research and suggests that referees use many criteria in their assessment. This finding, which refers to the second phase in Hirschauer's (2005, 2010) three-phase model of judgment (i.e., the judgment developed over the course of reading an application), is in line with results from studies that focused on the first phase (i.e., the expectations shared in a research community) and the third phase (i.e., the written post hoc rationalization of the judgment). More specifically, a systematic review identified a broad set of criteria peers refer to in the evaluation of grant applications (Hug & Aeschbach, 2020), and empirical studies demonstrated that scholars' notions of research quality and performance are multifaceted (Andersen, 2013; Bazeley, 2010; Gulbrandsen, 2000; Hemlin, 1993; Hemlin & Montgomery, 1990; Hug et al., 2013; Margherita et al., 2022; Mårtensson et al., 2016; Ochsner et al., 2013; Prpić & Šuljok, 2009). Moreover, the finding is in line with Hren et al. (2022) who showed that almost all of the 29 themes (or criteria) identified in the evaluation summaries written by grant panels are related to the panels' decisions.

The results of this study do not support one of the key findings of the research on fast-and-frugal heuristics, the less-is-more effect, which states that less information and computation can lead to more accurate judgments than more information and computation (Gigerenzer & Goldstein, 1996; Gigerenzer et al., 2022). In accordance with the less-is-more effect, Raab and Gigerenzer (2015) as well as Phillips et al. (2017) emphasize that fast-and-frugal trees can be as accurate as or more accurate than complex procedures. Empirical studies, however, show a more nuanced picture. Fast-and-frugal trees were found to be more accurate than (e.g., Dhami & Ayton, 2001; Wegwarth et al., 2009), as accurate as (e.g., Aikman et al., 2021; Jenny et al., 2013), or less accurate than (e.g., Woike et al., 2015) models integrating all cues. This range of results was also obtained by Phillips et al. (2017) who compared fast-and-frugal trees to six complex procedures using ten real-world datasets. They found that, in the testing set, the accuracy of fast-and-frugal trees (0.83) was higher than that of naïve Bayes (0.78) and CART (0.80), comparable to logistic regression (0.82), regularized logistic regression (0.83), and random forests (0.83), but lower than that of support-vector machines (0.86). Moreover, several studies demonstrated that when the size of the dataset for training was small (i.e., 15% of the total data), fast-and-frugal trees tended to do best, producing either comparable or slightly better results in the testing set than complex procedures (Laskey & Martignon, 2014; Martignon et al., 2008, 2012; Woike et al., 2017). When 50% of the data was included in the training set, logistic regression was 2–3% more accurate than fast-and-frugal trees (Martignon et al., 2008, 2012), which corresponds to the design and the results of the current study. The results of the present study

are thus in line with previous studies on fast-and-frugal trees but do not support the less-is-more effect.

The study has the following main limitations. First, it focused on a funding scheme for doctoral students and postdocs, and it analyzed data from a relatively small pool of referees. It thus remains to be investigated whether the findings can be generalized to a larger population of referees, other funding schemes, and other types of academic peer review. Second, the study examined the performance of only two decision rules (fast-and-frugal trees, logistic regression) and thus may not have considered the rule that can most accurately describe referees' judgments. Future studies may therefore include a broader range of heuristics and complex procedures. Third, the analysis focused on evaluation criteria provided by the funding organization and did not include characteristics of the applications, applicants, or referees (e.g., gender, age, discipline). The effect of potential bias factors has thus not been assessed. Future studies should therefore include bias factors, such as gender (Cruz-Castro & Sanz-Menendez, 2021; Sato et al., 2021; Schmaling & Gallo, 2023; Squazzoni et al., 2021), when examining the two judgment styles. Fourth, the referees' ratings were analyzed using a nomothetic, group-level approach (Beltz et al., 2016; Piccirillo & Rodebaugh, 2019), which may have favored the uniform judgment style. Future research may therefore employ an idiographic approach and examine the judgments of each referee individually. Fifth, the ratings of the proposals were treated as events independent of the referees. As grant peer review is inherently multilevel (Erosheva et al., 2020), future research may use multilevel approaches to examine the two judgment styles proposed in this study. Lastly, the study was based on the assumption that criteria scores reported on rating forms allow inferences to be made about the referees' evaluation process and the two judgments styles. However, data collected during the assessment process may be more appropriate for examining the two styles. For example, Vallée-Tourangeau et al. (2022) used the think aloud method to explore factors influencing the evaluation of grant applications.

The results of this study, if generalizable, have several implications for peer review research and practice. First, the results support the practice of many funding agencies to score applications on a small number of evaluation criteria. According to Langfeldt and Scordato (2016), funding agencies use few criteria for reasons of simplicity, clarity, flexibility, and efficiency. The present study demonstrated that as few as two evaluation criteria can explain the variability of the referees' overall judgments to a large extent, whereas including significantly more criteria improved the explained variation only slightly. This suggests that funding agencies using few criteria sacrifice little quantitative information in exchange for greater simplicity, clarity, and flexibility. Nevertheless, funders must be aware that referees use many evaluation criteria. To complement the quantitative criteria in the rating forms, it is therefore reasonable to retain the written assessments widely used by funding agencies. Moreover, funding agencies that do not ask referees to provide an overall score but calculate overall scores from a small number of criteria or agencies that base their funding decisions solely on criteria scores need to be aware that they ignore some of the evaluative information referees can provide and thus likely alter the results of the peer review process. Second, for modeling peer review judgments, fast-and-frugal trees can represent an alternative to more complex statistical methods, depending on the research context. When only few data are available or collecting data is costly, fast-and-frugal trees provide an effective alternative because trees are relatively robust against overfitting (Phillips et al., 2017) and produce comparable or slightly better results than complex procedures (Laskey & Martignon, 2014; Martignon et al., 2008, 2012; Woike et al., 2017). In addition, fast-and-frugal trees have several advantages that can outweigh the slight loss

in accuracy found in this and other studies. Trees typically use very little information (i.e., cues); they are computationally inexpensive; and they enable fast decisions. They are easy to understand, communicate, and apply because they are simple and transparent (Phillips et al., 2017). And trees, like other fast-and-frugal heuristics, are designed for studying decision-making situations characterized by uncertainty (Gigerenzer et al., 2022); peer review represents such a situation. Third, the present study has shown that referees mostly use the same criteria and integrate the criteria in the same way. This suggests that the disagreement effect (i.e., the low inter-rater reliability consistently observed in peer review) is less likely to be caused by referees using different criteria and integration rules.

As this study presented evidence that supports the uniform judgment style, future studies should attempt to falsify this judgment style. Future studies should also examine how referees' judgments can be characterized in those cases where the uniform style fails to provide an accurate and proper description. Lastly, and more broadly, peer review research should study the evaluation process of referees in more empirical and theoretical detail. This would advance our understanding of how peer review works and how we can develop it further.

Appendix

See Table 4 and Fig. 5.

Table 4 Verbal description of fast-and-frugal-trees describing referees' overall judgment with four, five, and six cues

No. of cues	Description
Four cues	Like the three-cue tree (Fig. 2c) but replace its lowest level by: If letter of recommendation < = 4, decide "not outstanding"; otherwise, assess career plan. If career plan < = 4, decide "not outstanding"; otherwise, decide "outstanding"
Five cues	Like the four-cue tree but replace its lowest level by: If career plan > 4, decide "outstanding"; otherwise, assess cooperation/network. If cooperation/network < = 4, decide "not outstanding"; otherwise, decide "outstanding"
Six cues	Like the five-cue tree but replace its lowest level by: If cooperation/network < = 4, decide "not outstanding"; otherwise, assess clarity of project goal. If clarity of project goal < = 4, decide "not outstanding"; otherwise, decide "outstanding"

Decision rule	Training set					Testing set				
	#Frug	%Frug	Acc	Sens	Spec	#Frug	%Frug	Acc	Sens	Spec
Fast-and-frugal trees										
1 cue	1	92%	.82	.90	.75	1	92%	.83	.87	.79
2 cues	1.6	88%	.86	.79	.92	1.5	88%	.87	.78	.97
3 cues	1.7	87%	.85	.88	.82	1.7	87%	.85	.84	.86
4 cues	1.8	86%	.86	.82	.90	1.8	86%	.87	.81	.94
5 cues	1.9	85%	.87	.88	.85	1.8	86%	.86	.81	.90
6 cues	1.8	86%	.87	.80	.92	1.8	86%	.87	.78	.96
Logistic regression										
2 cues (log-odds)	2	85%	.87	.96	.79	2	85%	.84	.88	.80
2 cues (cue validity)	2	85%	.87	.81	.92	2	85%	.87	.80	.95
All cues	13	0%	.94	.95	.92	13	0%	.89	.86	.92

Note: #Frug = absolute frugality; %Frug = relative frugality; Acc = accuracy; Sens = sensitivity; Spec = specificity.

Fig. 5 Performance of fast-and-frugal trees and logistic regressions in training (238 judgments) and testing (236 judgments). Stratified random sampling was used to create the two sets (each set: 50% judgments from the first referees, 50% judgments from the second referees)

Acknowledgements I am grateful to Anna Ekert-Centowska for her thoughtful comments and suggestions on earlier versions of this paper. I would like to extend my sincere thanks to Rüdiger Mutz for helpful discussions. I also wish to thank the reviewers for thoroughly engaging with the manuscript and for providing constructive feedback.

Funding Open access funding provided by University of Zurich.

Data availability Data is not available due to legal restrictions.

Declarations

Competing interest The author has no competing interests.

References

Aikman, D., Galesic, M., Gigerenzer, G., Kapadia, S., Katsikopoulos, K., Kothiyal, A., Murphy, E., & Neumann, T. (2021). Taking uncertainty seriously: Simplicity versus complexity in financial regulation. *Industrial and Corporate Change, 30*(2), 317–345. https://doi.org/10.1093/icc/dtaa024

Andersen, J. P. (2013). *Conceptualising research quality in medicine for evaluative bibliometrics*. University of Copenhagen.

Artinger, F. M., Gigerenzer, G., & Jacobs, P. (2022). Satisficing: Integrating two traditions. *Journal of Economic Literature, 60*(2), 598–635. https://doi.org/10.1257/jel.20201396

Arvan, M., Bright, L. K., & Heesen, R. (2022). Jury theorems for peer review. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/719117

Baldwin, M. (2020). Peer review. *Encyclopedia of the History of Science*. https://doi.org/10.34758/srde-jw27

Batagelj, V., Ferligoj, A., & Squazzoni, F. (2017). The emergence of a field: A network analysis of research on peer review. *Scientometrics, 113*(1), 503–532. https://doi.org/10.1007/s11192-017-2522-8

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education, 35*(8), 889–903. https://doi.org/10.1080/03075070903348404

Beckstead, J. W. (2007). A note on determining the number of cues used in judgment analysis studies: The issue of type II error. *Judgment and Decision Making, 2*(5), 317–325. https://doi.org/10.1017/S1930297500000632

Beltz, A. M., Wright, A. G. C., Sprague, B. N., & Molenaar, P. C. M. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment, 23*(4), 447–458. https://doi.org/10.1177/1073191116648209

Bornmann, L. (2008). Scientific peer review: An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture: Journal of the Sociology of Self-Knowledge, VI*, 23–38.

Bornmann, L. (2015). Complex tasks and simple solutions: The use of heuristics in the evaluation of research. *Journal of the Association for Information Science and Technology, 66*(8), 1738–1739. https://doi.org/10.1002/asi.23318

Bornmann, L., Ganser, C., & Tekles, A. (2022). Simulation of the h index use at university departments within the bibliometrics-based heuristics framework: Can the indicator be used to compare individual researchers? *Journal of Informetrics, 16*(1), 101237. https://doi.org/10.1016/j.joi.2021.101237

Bornmann, L., & Hug, S. E. (2020). Bibliometrics-based heuristics: What is their definition and how can they be studied? *Profesional De La Información, 29*(4), e290420. https://doi.org/10.3145/epi.2020.jul.20

Bornmann, L., & Marewski, J. N. (2019). Heuristics as conceptual lens for understanding and studying the usage of bibliometrics in research evaluation. *Scientometrics, 120*(2), 419–459. https://doi.org/10.1007/s11192-019-03018-x

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science*. State University of New York Press.

CoARA. (2022). *Agreement on reforming research assessment*. Retrieved from https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

Cruz-Castro, L., & Sanz-Menendez, L. (2021). What should be rewarded? Gender and evaluation criteria for tenure and promotion. *Journal of Informetrics, 15*(3), 101196. https://doi.org/10.1016/j.joi.2021.101196

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 97–118). Oxford University Press.

de Abreu Batista Júnior, A., Gouveia, F. C., & Mena-Chalco, J. P. (2021). Identification of promising researchers through fast-and-frugal heuristics. In Y. Manolopoulos & T. Vergoulis (Eds.), *Predicting the dynamics of research impact* (pp. 195–207). Springer.

Devyatkin, D., Suvorov, R., Tikhomirov, I., & Grigoriev, O. (2018). Scientific research funding criteria: An empirical study of peer review and scientometrics. In V. Sgurev, V. Jotsov, & J. Kacprzyk (Eds.), *Practical issues of intelligent innovations* (pp. 277–292). Springer. https://doi.org/10.1007/978-3-319-78437-3_12

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making, 14*(2), 141–168. https://doi.org/10.1002/bdm.371

Eblen, M. K., Wagner, R. M., RoyChowdhury, D., Patel, K. C., & Pearson, K. (2016). How criterion scores predict the overall impact score and funding outcomes for National Institutes of Health peer-reviewed applications. *PLoS ONE, 11*(6), e0155060. https://doi.org/10.1371/journal.pone.0155060

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin, 73*(3), 221–230. https://doi.org/10.1037/h0028695

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., & Lee, C. J. (2020). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances, 6*(23), eaaz4868. https://doi.org/10.1126/sciadv.aaz4868

European Commission. (2021). *Towards a reform of the research assessment system: Scoping report*. https://doi.org/10.2777/707440

Forscher, P. S., Brauer, M., Cox, W. T. L., & Devine, P. G. (2019). How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals? *PsyArxiv*. https://doi.org/10.31234/osf.io/483zj

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE.
Gallo, S. A., Sullivan, J. H., & Glisson, S. R. (2016). The influence of peer reviewer expertise on the evaluation of research funding applications. *PLoS ONE, 11*(10), e0165147. https://doi.org/10.1371/journal.pone.0165147
Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.
Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650
Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.
Gigerenzer, G., Reb, J., & Luan, S. (2022). Smart heuristics for individuals, teams, and organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 9*(1), 171–198. https://doi.org/10.1146/annurev-orgpsych-012420-090506
Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.
Gläser, J., & Laudel, G. (2005). Advantages and dangers of 'remote' peer evaluation. *Research Evaluation, 14*(3), 186–198. https://doi.org/10.3152/147154405781776085
Grimaldo, F., Marušić, A., & Squazzoni, F. (2018). Fragments of peer review: A quantitative analysis of the literature (1969–2015). *PLoS ONE, 13*(2), 14. https://doi.org/10.1371/journal.pone.0193148
Grove, W. M. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of Clinical Psychology, 61*(10), 1233–1243. https://doi.org/10.1002/jclp.20179
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293–323.
Gulbrandsen, M. J. (2000). *Research quality and organisational factors: An investigation of the relationship*. Norwegian University of Science and Technology.
Hamann, J., & Beljean, S. (2017). Academic evaluation in higher education. In J. C. Shin & P. N. Teixeira (Eds.), *Encyclopedia of International Higher Education Systems and Institutions* (pp. 1–7). Springer. https://doi.org/10.1007/978-94-017-9553-1_295-1
Heesen, R. (2019). The necessity of commensuration bias in grant peer review. *PhilSci-Archive*. http://philsci-archive.pitt.edu/id/eprint/15930
Hemlin, S. (1993). Scientific quality in the eyes of the scientist: A questionnaire study. *Scientometrics, 27*(1), 3–18. https://doi.org/10.1007/bf02017752
Hemlin, S., & Montgomery, H. (1990). Scientists' conceptions of scientific quality: An interview study. *Science Studies, 3*(1), 73–81.
Hirschauer, S. (2004). Peer Review Verfahren auf dem Prüfstand. Zum Soziologiedefizit der Wissenschaftsevaluation. *Zeitschrift Fur Soziologie*, *33*(1), 62–83.
Hirschauer, S. (2019). Urteilen unter Beobachtung: performative publizität im peer review. In S. Nicolae, M. Endress, O. Berli, & D. Bischur (Eds.), *(Be)Werten: Beiträge zur sozialen Konstruktion von Wertigkeit* (pp. 275–298). Springer. https://doi.org/10.1007/978-3-658-21763-1
Hirschauer, S. (2005). Publizierte Fachurteile: Lektüre und Bewertungspraxis im peer review. *Soziale Systeme, 11*(1), 52–82. https://doi.org/10.1515/9783110511369-004
Hirschauer, S. (2010). Editorial judgments: A praxeology of "voting" in peer review. *Social Studies of Science, 40*(1), 71–103. https://doi.org/10.1177/0306312709335405
Hirschauer, S. (2015). How editors decide: Oral communication in journal peer review. *Human Studies, 38*(1), 37–55. https://doi.org/10.1007/s10746-014-9329-x
Hren, D., Pina, D. G., Norman, C. R., & Marušić, A. (2022). What makes or breaks competitive research proposals? A mixed-methods analysis of research grant evaluation reports. *Journal of Informetrics, 16*(2), 101289. https://doi.org/10.1016/j.joi.2022.101289
Hug, S. E. (2022). Towards theorizing peer review. *Quantitative Science Studies, 3*(3), 815–831. https://doi.org/10.1162/qss_a_00195
Hug, S. E., & Aeschbach, M. (2020). Criteria for assessing grant applications: A systematic review. *Palgrave Communications*, *6*(37). https://doi.org/10.1057/s41599-020-0412-9
Hug, S. E., & Ochsner, M. (2022). Do peers share the same criteria for assessing grant applications? *Research Evaluation, 31*(1), 104–117. https://doi.org/10.1093/reseval/rvab034
Hug, S. E., Ochsner, M., & Daniel, H. D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383. https://doi.org/10.1093/reseval/rvt008

Jenny, M. A., Pachur, T., Lloyd Williams, S., Becker, E., & Margraf, J. (2013). Simple rules for detecting depression. *Journal of Applied Research in Memory and Cognition, 2*(3), 149–157. https://doi.org/10.1016/j.jarmac.2013.06.001

Johnson, D. R., & Hermanowicz, J. C. (2017). Peer review: From 'sacred ideals' to 'profane realities.' In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (Vol. 32, pp. 485–527). Springer. https://doi.org/10.1007/978-3-319-48983-4

Johnston, R., Jones, K., & Manley, D. (2018). Confounding and collinearity in regression analysis: A cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & Quantity, 52*(4), 1957–1976. https://doi.org/10.1007/s11135-017-0584-6

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Kang, D., Ammar, W., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)* (pp. 1647–1661). https://doi.org/10.18653/v1/N18-1149

Langfeldt, L., & Scordato, L. (2016). *Efficiency and flexibility in research funding: A comparative study of funding instruments and review criteria*. Nordic Institute for Studies in Innovation, Research and Education.

Laskey, K., & Martignon, L. (2014). Comparing fast and frugal trees and Bayesian networks for risk assessment. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the ninth international conference on teaching statistics*. International Statistical Institute.

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science, 82*(5), 1272–1283. https://doi.org/10.1086/683652

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2–17. https://doi.org/10.1002/asi.22784

Lindner, M. D., Vancea, A., Chen, M.-C., & Chacko, G. (2016). NIH peer review: Scored review criteria and overall impact. *American Journal of Evaluation, 37*(2), 238–249. https://doi.org/10.1177/1098214015582049

Margherita, A., Elia, G., & Petti, C. (2022). What is quality in research? Building a framework of design, process and impact attributes and evaluation perspectives. *Sustainability, 14*(5), 3034. https://doi.org/10.3390/su14053034

Mårtensson, P., Fors, U., Wallin, S. B., Zander, U., & Nilsson, G. H. (2016). Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Research Policy, 45*(3), 593–603. https://doi.org/10.1016/j.respol.2015.11.009

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 119–140). Oxford University Press.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2012). Naïve, fast, and frugal trees for classification. In P. M. Todd, G. Gigerenzer, & ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 360–378). Oxford University Press.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology, 52*(6), 352–361. https://doi.org/10.1016/j.jmp.2008.04.003

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 189–211). John Wiley and Sons.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Ochsner, M., Hug, S. E., & Daniel, H. D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(2), 79–92. https://doi.org/10.1093/reseval/rvs039

Phillips, N. D., Neth, H., Woike, J. K., & Gaissmaier, W. (2017). FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making, 12*(4), 344–368.

Piccirillo, M. L., & Rodebaugh, T. L. (2019). Foundations of idiographic methods in psychology and applications for psychotherapy. *Clinical Psychology Review, 71*, 90–100. https://doi.org/10.1016/j.cpr.2019.01.002

Pina, D. G., Buljan, I., Hren, D., & Marušić, A. (2021). A retrospective analysis of the peer review of more than 75,000 Marie Curie proposals between 2007 and 2018. *eLife, 10*, e59338. https://doi.org/10.7554/eLife.59338

Porter, A. L., & Rossini, F. A. (1985). Peer review of interdisciplinary research proposals. *Science Technology & Human Values, 10*(3), 33–38. https://doi.org/10.1177/016224398501000304
Prabhakaran, S. (2016). *Information value: Performance analysis and companion functions for binary classification models*. Retrieved from http://r-statistics.co/Information-Value-With-R.html
Prpić, K., & Šuljok, A. (2009). How do scientists perceive scientific quality. In K. Prpić (Ed.), *Beyond the myths about the natural and social sciences: A sociological view* (pp. 205–245). Institute for Social Research.
R Core Team. (2021). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org
Raab, M., & Gigerenzer, G. (2015). The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.01672
Reinhart, M. (2009). Peer review of grant applications in biology and medicine: Reliability, fairness, and validity. *Scientometrics, 81*(3), 789–809. https://doi.org/10.1007/s11192-008-2220-7
Reinhart, M., & Schendzielorz, C. (2021). Peer review procedures as practice, decision, and governance: Preliminaries to theories of peer review. *SocArXiv*. https://doi.org/10.31235/osf.io/ybp25
Rockey, S. (2011). *Correlation between overall impact scores and criterion Scores*. Retrieved 25 Oct 2022, from https://nexus.od.nih.gov/all/2011/03/08/overall-impact-and-criterion-scores/
Sato, S., Gygax, P. M., Randall, J., & Schmid Mast, M. (2021). The leaky pipeline in research grant peer review and funding decisions: Challenges and future directions. *Higher Education, 82*(1), 145–162. https://doi.org/10.1007/s10734-020-00626-y
Scarr, S. (1982). Anosmic peer review: A rose by another name is evidently not a rose. *Behavioral and Brain Sciences, 5*(2), 237–238.
Schmaling, K. B., & Gallo, S. A. (2023). Gender differences in peer reviewed grant applications, awards, and amounts: A systematic review and meta-analysis. *Research Integrity and Peer Review, 8*(1), 2. https://doi.org/10.1186/s41073-023-00127-3
Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulleting, 134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207
Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P., & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances, 7*(2), eabd0299. https://doi.org/10.1126/sciadv.abd0299
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
Vallée-Tourangeau, G., Wheelock, A., Vandrevala, T., & Harries, P. (2022). Peer reviewers' dilemmas: A qualitative exploration of decisional conflict in the evaluation of grant applications in the medical humanities and social sciences. *Humanities and Social Sciences Communications, 9*(1), 70. https://doi.org/10.1057/s41599-022-01050-6
Wang, Y., Luan, S., & Gigerenzer, G. (2022). Modeling fast-and-frugal heuristics. *PsyCh Journal, 11*(4), 600–611. https://doi.org/10.1002/pchj.576
Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2009). Smart strategies for doctors and doctors-in-training: Heuristics in medicine. *Medical Education, 43*(8), 721–728. https://doi.org/10.1111/j.1365-2923.2009.03359.x
Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision, 4*, 234–260. https://doi.org/10.1037/dec0000086
Woike, J. K., Hoffrage, U., & Petty, J. S. (2015). Picking profitable investments: The success of equal weighting in simulated venture capitalist decision making. *Special Issue on Simple versus Complex Forecasting, 68*(8), 1705–1716. https://doi.org/10.1016/j.jbusres.2015.03.030
Würth, S., Milzow, K., & Egger, M. (2017). *Influence of evaluation criteria on overall assessment in peer review of project grants submitted to the Swiss National Science Foundation*. Eighth International Congress on Peer Review and Scientific Publication, Chicago. Retrieved from https://peerreviewcongress.org/abstract/influence-of-evaluation-criteria-on-overall-assessment-in-peer-review-of-project-grants-submitted-to-the-swiss-national-science-foundation/
Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva, 9*(1), 66–100. https://doi.org/10.1007/BF01553188