# Quis judicabit ipsos judices? A case study on the dynamics of competitive funding panel evaluations

## Abstract

Securing research funding is essential for all researchers. The standard evaluation method for competitive grants is through evaluation by a panel of experts. However, the literature notes that peer review has inherent flaws and is subject to biases which can arise from differing interpretations of the criteria, the impossibility for a group of reviewers to be experts in all possible topics within their field, and the role of affect. As such, understanding the dynamics at play during panel evaluations is crucial to allow researchers a better chance at securing funding, and also for the reviewers themselves to be aware of the cognitive mechanisms underlying their decision-making. In this study, we conduct a case study based on application and evaluation data for two social sciences panels in a competitive state-funded call in Portugal. Using a mixed-methods approach, we find that qualitative evaluations largely resonate with the evaluation criteria, and the candidate's scientific output is partially aligned with the qualitative evaluations, but scientometric indicators alone do not significantly influence the candidate's evaluation. However, the polarity of the qualitative evaluation has a positive influence on the candidate's evaluation. This paradox is discussed as possibly resulting from the occurrence of a halo effect in the panel's judgment of the candidates. By providing a multi-methods approach, this study aims to provide insights which can be useful for all stakeholders involved in competitive funding evaluations.

## Introduction

Applying for competitive funding is an inherent part of the researcher profession, and something which virtually all researchers will do at some point. The increasingly competitive nature of the profession is becoming a more prominent topic with far-reaching policy, economic, and societal implications, largely due to the emergence of the now well-known "publish or perish" dynamics (Backes-Gellner & Schlinghoff, 2010; McGrail et al., 2006). Although this is still an ongoing debate, it has been noted in the literature that part of this can be due to evaluation schemes which might have created perverse incentives to gamify the system (Martin, 2011; Stephan, 2012), leading to maximization of indicators with the sole goal of securing funding and not necessarily leading to good science (Young, 2015); indeed, and although it is unlikely that this is the sole cause, the rates of worldwide innovation have been steadily decreasing (Huebner, 2005). Simultaneously, it has been shown that funding tends to be concentrated in "scientific elites" (Larivière et al., 2010), causing further issues as it has also been shown that such concentration of resources tends to yield diminishing returns (Mongeon et al., 2016). Attempts to solve the issues of exploitation of the system have taken the form of funding lotteries (Smaldino et al., 2019) – the rationale being that applications for competitive funding should only need to pass a basic screening for quality, and after reaching a threshold for admissibility, they are simply drawn at random – thus preventing indicator gamification, while simultaneously allowing a "fighting chance" for projects on topics which are methodologically sound, but would have lower chances of receiving funding when compared to "hot topics".

Academic employment has become precarious (Lempiäinen, 2015); lack of institutional funding, which by itself tends to also be in part competitive, pushes researchers into securing their own funding (Laudel, 2006). This can sometimes be necessary to secure their own employment, i.e., through a research grant, fellowship, or a funded project. This is the case for Portugal; although there exists a legal structure for contracting career researchers on a permanent basis, a substantial number of researchers either hold teaching contracts, or are hired by specific research projects and not necessarily by the institution *per se*. As such, researchers rely, for the most part, on state grants and state funded contracts in other to secure employment. Such is the case of the Individual Call to Scientific Employment Stimulus – a program first opened in 2018, promoted by the national research agency – Fundação para a Ciência e Tecnologia - with the goal of replacing post-doctoral grants with six-year contracts. This is highly desirable by researchers, not only due to the added longevity of the contract when compared to post-doctoral grants, but also because of the increased income and job security provided by a work contract when compared to simply being a grant-holder; this shift was met with some initial resistance by institutions due to the added costs in taxation for employing a researcher by contract, as opposed to a grant-holder whose income is tax-free in Portugal. Nevertheless, this call has been ongoing since then, and funds several hundred researchers at varying stages every year, mostly early career; however, this is out of several thousand candidates. As such, the difference between receiving funding or not is razor thin – because of this, understanding the inner workings of panel evaluations is invaluable for researchers trying to gain a foothold in the research arena.

The *de facto* practice for competitive funding evaluation is through a panel of independent peers. Peer review is an important part of the scientific endeavor, serving as a way of ensuring quality control and validity of findings in scientific publications (Alberts et al., 2008). However, in the context of a panel of juries, the dynamics differ substantially from the standard "editor plus two referees" format commonly seen in journals. First, the number of peers tends to be higher; and second, they interact directly with one another, generally with the goal of attaining consensus, rather than providing independent and individual assessments (Bozeman, 1993). However, there are limitations to this approach due to the inherent subjectivity in the process of translating evaluation criteria into a classification (Zhu et al., 2022). This exacerbated by two aspects: first, no expert is all-knowing, and modern research careers and agendas are largely multidisciplinary causing situations where an application can fall outside of the expertise of the panel members (Zhu et al., 2020, 2022); and second, the understanding and interpretation of the evaluation guidelines can be vague or imprecise, causing hesitance on ascribing a particular rating to a candidate (Zhu et al., 2022). Additionally, there is a known documented disagreement effect in peer-review, referring to a low inter-rater reliability (Hug & Ochsner, 2022) – that is, differing reviewers can have radically different opinions on the same application. Indeed, because of the aforementioned, the literature has shown that the exact same grant application can be approved or rejected solely on the basis of which reviewer was assigned to it (Pier et al., 2018). More recent studies deemed these evaluations unreliable with inter-reviewer score correlations of 0.2 (Jerrim & Vries, 2020), while other authors go even further by noting that there is a great deal of chance involved in this process (Roumbanis, 2021). As these reviewing dynamics shape and influence the careers of academics all over the world, opening the black box in which they currently operate is essential.

Several works have explored the dynamics of panel evaluation. One such work is Lamont's book *How professors think* (2009), which focuses on the mechanics of panel evaluation of competitive grants in the social sciences. Lamont provides compelling evidence on the underlying complexity of the decision-making in such a context; panelists go beyond the formal criteria for evaluation and also apply informal, or "evanescent" criteria – such as perceived intelligence, elegance, personal, and moral qualities of the candidate – when deliberating on their ratings. Indeed, other classic works have portrayed academics as striving to acquire intellectual capital (Bourdieu, 1999), through which they can impose their vision on the academic landscape (Bourdieu, 1988), whereas others have noted how reputation and prestige can "taint" decisions through an halo effect (Merton, 1996). This study, however, does not aim at entering the debate on whether or not subjectivity has a place in panel evaluations. Rather, it aims to create a better understanding of how such complex decision-making takes place, by going beyond the explicit criteria for evaluations and determining the role affect plays in ascribing scores to candidates.

This study presents itself as a case study of panel evaluation dynamics using application and evaluation data from the 3rd (2020) and 4th (2021) edition of the Individual Call to Scientific Employment Stimulus. It employs a mixed-methods approach in order to gain insights into how individual applicants are selected by panelists in a competitive funding call. Furthermore, it aims to demonstrate and propose several methodologies which can be employed to judge the scientometric validity of panel evaluations. The paper will begin with the presentation of the methods employed, followed by the three primary analysis. It will then conclude with a summary of findings and a discussion on implications, limitations, and future directions for this line of research.

## Method

### *Data collection*

The data for this study was collected from two primary sources. The first were the panel evaluations for the 3rd (2020) and 4th (2021) edition of the Individual Call to Scientific Employment Stimulus, promoted by the Fundação para a Ciência e Tecnologia (FCT) in Portugal. Although there were various panels, corresponding to a multitude of fields of science, only one panel was used for each edition due to limited data availability. Due to privacy concerns, it will not be stated which specific panels the data was collected for, but we note that these were within the social sciences, and thus the findings might not be applicable to other fields of science. Additionally, the data refers to candidate at the Junior level – up to five years of experience following PhD conclusion. The panel results included both quantitative and qualitative evaluations of both the candidate and the project. This data was collected for analysis, although it must be noted that the data on project evaluation was not analyzed for two reasons – first, it only corresponds to a smaller fraction of the final score (30%); and second, project evaluation is highly subjective and not adequate for the type of analysis which was planned for this exercise. A total of 144 candidates were identified, corresponding to 73 for the 3rd edition and 71 for the 4th edition of the call.

Following this, scientometric data was obtained from the Portuguese scientific curriculum management website – Ciência Vitae. The reason for using this specific source is that it is the official curriculum platform which is used in this call, thus representing the actual information which the evaluation panels had access to. Since these calls were conducted in differing time periods, the content of the curriculums had, necessarily, shifted over time. Thus, for data collection, indicators were only considered up to the year in which the call occurred; for the 3rd edition, up to 2020; and for the 4th edition, up to 2021, thus providing a snapshot of what the curriculum would have looked like at time of submission. Unfortunately, the calls closed in February of the respective years, and the indicators in the platform did not have a month timestamp, which means that there is a “blind spot” of roughly two months during which indicators might not have been counted, an unavoidable limitation due to the nature of the data. Additionally, some candidates had either closed their Ciência Vitae profile at the time of data collection or set its visibility to private. As such, these 16 candidates were not considered for the analysis which relied on scientometric data, leaving a sample size of 128 for those specific analysis.

Ciência Vitae profiles are broadly organized into six main categories: Education, Employment, Production, Activities, Prizes, and Projects. Each of these sections has sub-sections. Data was collected for each of these sub-sections, but aggregation had to be subsequently done in order to keep the number of variables at a manageable level, as will be described further ahead in the manuscript.

### *Panel composition and operation*

In accordance with the regulations for the Individual Call to Scientific Employment Stimulus program (FCT, 2020), the evaluation panels are nominated following deliberation by FCT's board of directors. Furthermore, they are to be comprised, preferably, by international experts of known merit. Additionally, the panels can request assistance from external evaluators if necessary. The panels are required to apply each of the evaluation parameters to the submitted applications, write a report on each of them, order them by score, identify which ones are eligible for funding, and compile a final report which includes the results, and also feedback on the evaluation system. Although the regulations do not mention how many members each panel should have, we note that the 3rd edition panel was comprised of 11 individuals, and the 4th edition panel had 14 members.

Proceedings must be written for each panel meeting, with a summary of what was discussed, which members of the panel were present, the rationale for the evaluations, and whatever other issues were discussed. Although the regulations suggest that the panel has the final verdict on which candidates are funded, this is not necessarily the case. Consulting the proceedings confirms that the panels ascribe a final score to each candidate, but do not decide what is the cutoff point in the ordered lists. Thus, when they submit the final report for their respective panel, they are unaware of which candidates will in fact secure funding.

Instead, after the disciplinary panel meetings are done, the chairs of each panel hold a general meeting. It is this "Coordinating Evaluation Panel" – comprised by 26 individuals – which decides how the available positions will be distributed by each disciplinary panel and by level of application. Thus, the success rate is defined based on available funding and number of candidates; this was set at 8.2% for the 3rd edition, and 10.7% for the 4th edition. As such, it is only at this point that the cutoff point is decided and applied to the ordered listings of candidates in each panel.

Following this, provisional results are published, and candidates have access to the full applications, evaluations, and scores of every applicant in their own panel. Candidates who are rejected can file an appeal, which will be evaluated by the same panel which rated the original application. Only after this second round of evaluations are the results deemed final. However, if the application is again rejected, candidates still have available one final option for appeal – they can escalate to a formal complaint, which is then evaluated by a separate panel of experts. Following this, the final results can be amended or not based on their decision, and there are no further options for appeal.

### *Variables*

The data extracted from the Ciência Vitae platform was organized in a way to preserve its original structure in the platform; however, data reduction was necessary due to the small sample size. Literature suggests a minimum of two participants per variable for regression analysis (Austin & Steyerberg, 2015). Considering this, categories which were underrepresented in the candidates' curriculums were merged; categories which had an eminent positioning in the qualitative analysis were kept as stand-alone for counting purposes. The organization scheme is as such:

*Activities* refers to various activities conducted by the candidate, such as belonging to associations, committees, having participated in conferences, organized events, conducted peer review, teaching activities, and other non-academic activities. *Production* refers to the scientific output of the candidate, including books, book chapters, conference production, cultural/artistic production, published papers, media production, thesis, and other production. *Projects* refers to competitive projects or funding, such as grants, competitive projects, and other type of projects. *Work* is the professional path of the candidate, including number of scientific jobs, teaching jobs, and other types of jobs. Finally, *prizes* refers to titles, awards, and other distinctions obtained by the candidate. The aggregation scheme for the various categories is shown in Table 1:

<INSERT TABLE 1 HERE>

Based on the results from Analysis 1 – the content analysis, which will be presented further ahead – it became evident that multiple counting methods were necessary for an accurate representation of the panel's evaluation process, since there was indication that authorship and internalization played an important role. As such, each of these scientometric indicators were counted in five different ways. In *raw counting*, they were counted simply as they are – one item, one count. In *international counting*, the item was only counted if it was deemed an internationalized item. The criterion for establishing internationalization was whether or not the item was in the English language. Lacking data on the localization of every individual item, this was deemed a satisfactory compromise – English is, after all, the *lingua franca* of science (Sano, 2002) and it is reasonable to assume that materials written in other languages tend to have a regional, rather than an international scope. Additionally, in Portugal there are policy-related motives to understand internationalization as English-based. First, one of the key criteria for evaluation of R&D units in Portugal is their degree of internationalization as measured by number of publications in high-impact international journals (Horta, 2008), the vast majority of which are in English (Mueller et al., 2006; Seglen, 1997); second, this is acknowledged by institutional stakeholders, who encourage researchers to publish mainly in English (Pinto & Sá, 2020); and third, Portugal's primary international scientific partner, in terms of co-authored publications, is the United Kingdom (Patricio, 2010) – thus leading to a baseline bias towards publishing in English. *First author counting* only counts items where the candidate is the first author; whereas *Single author counting* only counts items where the candidate is the only author. Finally, *Fractional counting* weights the item by the number of co-authors – for example, an item with two authors is counted as 0.5, and an item with four authors is counted as 0.25. It is important to note that non-raw counting was not possible for all items since some of these categories do not have authorship or internationalization data; notably, authorship-based counting was only available for items under the Production category, and internationalization-based counting was only available for items under the Production and Activities categories.

Additionally, variables were computed to consider other aspects of internationalization, and other aspects which are, at the least, useful for control purposes. *Unique job countries Count* is a count variable indicating the number of unique countries in which the candidate held a position identified in the *Work* section of the Ciência Vitae profile; *Maternity* is a dummy variable indicating whether or not the candidate reported a maternity leave, which is important since this is stated in the evaluation guidelines as a criterion; reference category is having a maternity leave. *Gender* is a dummy variable indicating the participant's gender, with the reference category being males – this is also an important control variable due to the existence of a well-studied gender gap in the academia (Abramo et al., 2009; Frandsen et al., 2020; Santos et al., 2020). *PhD Uni Top Ranked* is a dummy variable indicating whether the candidate's PhD institution is in the top 500 of the Shanghai World University Ranking – important not only for control purposes due to the differing productivity dynamics of research-oriented institutions (Kwiek & Roszka, 2021), but also because this topic emerged several times in the content analysis. Reference category is being in a top-ranked institution. *Distinction* is a dummy variable indicating whether the candidate received a distinction, such as a *cum laude*, in his or her PhD degree, the reference category being having received such a distinction; this was included since it was mentioned by the panels, as noted in the content analysis. Finally, *Time since PhD* is the time, in years, elapsed since the conclusion of the candidate's PhD, also included both for control purposes – since productivity tends to accrue over time (Allison et al., 1982; Allison & Stewart, 1974) – but also since the PhD conclusion year was consistently mentioned in the qualitative evaluations.

*Procedure*

Three separate but inter-related analysis were conducted. In this section, we will provide a general overview of the analytical roadmap and the rationale behind it; for ease of reading, more specific details on each analysis' interpretation are provided in the respective section.

The first analysis is a content analysis (Drisko & Maschi, 2016) in which two corpora of materials were analyzed – the first were the evaluation guidelines for the call, and the second were the 144 qualitative evaluations provided to each application. The goal of this analysis was to identify both the explicit and implicit criteria for candidate evaluation. A complementary analysis was conducted – sentiment analysis. This methodology hails from the field of natural language processing, and aims to identify the affectual state of the writer behind a sentence by analyzing its polarity (Mohammad, 2016; Shaikh et al., 2007). The rationale behind this analysis is that criteria which implicitly elicit a stronger sentiment on the panelist will translate into polarity-loaded statements, rather than neutral ones. The polarity-loading of the evaluations was then used in subsequent analysis.

The second analysis is a classic scientometric analysis based on a set of OLS regressions (Hair et al., 2014; Montgomery et al., 2021), with five models, with the dependent variable being the candidate's quantitative evaluation score, and the independent variables being the various scientometric indicators described above as well as the control variables. Five models were specified; the dependent variable is always the same, but the type of counting varies by model, using the five different counting types which were described above. The goal of this analysis was to determine if scientometric indicators which were identified as relevant criteria in the previous analysis did in fact have an impact on the candidate's score.

The third analysis is a mixed-methods analysis which crosses the results from the qualitative analysis and the scientometric data. Individuals with positive, neutral, and negative polarity on various qualitative evaluations are compared based on the corresponding metric to determine whether there is a correspondence between the qualitative evaluation and the quantitative indicators – in other words, if an individual which was evaluated as publishing many papers does in fact have more papers than individuals with neutral or negative qualitative evaluations. This was done through ANOVA with Tukey's HSD post-hoc tests (Tukey, 1953). In addition, a polarity rating was computed for each candidate – based on the sum of negative (-1), neutral (0), and positive (1) occurrences, and used in a regression with the candidate score as the dependent variable, providing another measure of qualitative-quantitative correspondence.

**Results**

*Analysis 1 – Content analysis*

In this analysis, the primary goal was ascertaining what were the criteria – both explicit and implicit – for candidate evaluation. The explicit criteria were identified based on a simple reading of the call's regulations. The implicit criteria were identified based on a content analysis of the qualitative evaluation which was provided alongside the quantitative evaluation. The goal of this was to identify which aspects of the curriculums were more often mentioned by the panel; the rationale being that aspects mentioned more often are likely to have a larger impact on the final quantitative assessment.

Additionally, chi-square tests were conducted to identify whether the frequency of occurrence for each theme differed across panels.

The evaluation guide for the call is very similar for both editions[1], with only minor changes in the organization of the content. It begins with a list of four items stating that the merit of the candidate is based on his i) Curriculum Vitae, ii) the Motivation letter, iii) the CV Synopsis, and iv) if there is any interruption of scientific activity (this criterion is not explicitly stated in the criteria list for the 3rd edition but is mentioned in the following text nonetheless). The Curriculum Vitae refers to the candidate's curriculum, which must be uploaded through the national scientific curriculum management platform – Ciência Vitae – which is similar to ORCID, but specific to national state-funded calls. The motivation letter is self-explanatory and a common requirement for most job applications including non-academic ones. The CV synopsis is simply a summary of the CV written by the candidate where he or she indicates the highlights of the last five years. This is expected to duplicate the content of the main CV since at the Junior level the last five years might, realistically, represent the totality or majority of the applicant's CV – recall that this level of applications is exclusive to individuals who have completed their PhD less than five years prior to the call.

Following this list, the document states that the merit of the candidate is based on the aforementioned items:

> "(…) with emphasis on scientific, technological, cultural and/or artistic achievements and the applied research or research based in practice considered by the applicant as the most relevant or the most impactful. This criterion also considers other aspects highlighted by the applicant such as her/his internationalization, management of science, technology and innovation programmes or projects, scientific supervision, outreach activities and dissemination of knowledge, namely for the promotion of culture and scientific practices." (FCT, 2021, p. 6)

This first paragraph indicates several important aspects of evaluation: scientific and technological achievement, as well as cultural and artistic. The degree of internationalization, management activities, projects, supervisions, outreach, and dissemination activities. The guide continues as such:

> "The evaluation must consider the career level selected by the applicant, particularly in what concerns the evaluation of scientific independence (for principal and coordinating researchers) and scientific leadership (for coordinating researchers). An eventual mismatch of an application in relation to the researcher contract level may be penalized by the evaluation panel, e.g. an applicant applying for an Assistant researcher position but already having research seniority and scientific independence may be penalized." (FCT, 2021, p. 6)

This section highlights the consideration of career stage as an important aspect, albeit moreso for higher tiers of the call, of which this is not the case. It also indicates that applying at a level for which a participant is overqualified might result in disqualification. Finally:

[1] 3rd Edition:
https://www.fct.pt/apoios/contratacaodoutorados/empregocientifico/docs/CEECIND_3rd_Evaluation_Guide.pdf
4th Edition:
https://www.fct.pt/apoios/contratacaodoutorados/empregocientifico/docs/CEECIND_4th_Evaluation_Guide.pdf

> “Although taking into account the full professional path of the applicant, the evaluation should be focused on the last 5 years, with the following exceptions:
>
> - Junior researchers with less than 5 years of scientific activity;
>
> - Researchers who have interrupted their scientific activity due to maternity/paternity leave and/or serious illness, as well as other interruptions. In these situations, the evaluation should be focused in the last 5 or less working years of scientific activity, as described in the CV synopsis and explained by the applicant in the justification for the interruption of scientific activity. Please note that for the Junior Researcher level PhD holders for more than 5 years will be eligible if they had interruptions in their scientific activity due to maternity/paternity leave and/or serious illness.” (FCT, 2021, p. 6)

In this final section it states that the last five years should have greater weight on the evaluation. The data for this exercise was based on the Junior level – which have obtained the PhD less than five years prior to the call – and as such there is not an expectation of substantial career data beyond that point. Nevertheless, the full CV was fully considered for the quantitative analysis which will be reported in the following section.

Having these preliminary categories in mind, a content analysis was conducted for the full 144 qualitative evaluations. A first reading was conducted with several coding categories based on the aforementioned aspects which were explicitly referred in the evaluation guide. Throughout this first reading, *in-vivo* categories were also created based on non-explicit topics which emerged frequently (for example, authorship order) or which were considered noteworthy for standing out. After this first reading, the coding sheet was analyzed and refined to coalesce redundant categories, and a second reading was done for definitive coding. The coding sheet was as such:

<INSERT TABLE 2 HERE>

Applying this classification scheme, we computed the relative frequency of occurrence, globally and for each edition, which is as follows:

<INSERT TABLE 3 HERE>

The first noteworthy finding is the declared importance of publications, which was the most commonly emerging theme – the 4E Panel mentioned them in all evaluations, and the 3E Panel mentioned them in all but 3. This was followed by the date of the PhD’s conclusion, mentioned in 86.11% of all evaluations – however, this significantly differed by panel ($\chi^2$ (1) = 11.839, p < .001), with the 3E panel mentioning this more often (95.89%) than the 4E panel (76.06%). In third place is Scientific Work, mentioned in 74.31% of evaluations, and with no significant differences across panels. In fourth place is the degree of Internationalization, mentioned in 64.58% of evaluations, again with no notable differences across panels. Following this is conferences (63.19%), teaching (47.92%), projects (38.89%), book chapters (37.50%), and grants (36.81%). None of these differ across panels. The remaining themes were comparatively less common but can be seen in Table 3. Two noteworthy findings can be pointed out in these less common categories: one relates to first/single/co-authorship, which was mentioned twice as often (35.62%) in the 3E panel, when compared to the 4E panel (18.31%), a significant difference (χ$^2$ (1) = 5.459, p < .05). Additionally, educational grades were mentioned much more often in the 4E panel (23.94%) when compared to the 3E panel (5.48%), also a significant difference (χ$^2$ (1) = 9.851, p < .01). Overall, the themes which emerged in the qualitative evaluations largely resonated with the explicit evaluation criteria indicated in the evaluation guidelines.

A secondary analysis – sentiment analysis – was also conducted[2] (Table 4). For each thematic occurrence, it was classified based on its polarity – whether it was positive, negative, or neutral. For example, on the “Publications” theme, an occurrence was considered to have positive polarity if it framed the publication rate in a positive light (e.g., “the candidate has a substantial number of publications”), negative polarity if it reflected an insufficient productivity (e.g., “the candidate does not have many papers in the field”), or neutral if it did not carry any obvious polarity or consisted of a simple factual statement (e.g., “the candidate has published five papers”). To avoid potential semantic ambiguity, positive or negative polarity was only considered if the statement unambiguously exhibited positivity or negativity; for example, “the candidate has a few papers” could be interpreted in a variety of manners with regards to polarity, and as such was counted as neutral if it could not be contextually disambiguated; but “the candidate has few papers” would objectively indicate a lack of publications, and thus coded as having negative polarity. The polarity distribution is shown in the following table:

<INSERT TABLE 4 HERE>

This analysis highlights several key characteristics of the qualitative evaluation. First, most of the mentions of publication rates have some degree of polarity, with only 26.76% being neutral in nature. This suggests that publications should have some degree of influence on the quantitative grade, since they are deemed important enough to elicit an active polarity-loaded comment from the panel. Also notable, a substantial number of mentions of publication rates are indications of insufficiency of published works (40.85%). In opposition, the date of conclusion of the PhD, although being one of the most recurring themes, is always presented as a merely factual sentence. Our interpretation is that stating the PhD year is simply a confirmation of the candidate’s eligibility, since there is a 5-year cutoff as stated in the evaluation guide. Likewise, scientific work is mentioned frequently but in a largely neutral manner (71.96%); when polarity is present, it is more often portraying the candidate’s work as positive (21.5%) rather than negative (6.54%). More polarizing is the degree of internationalization; 25.81% mentions carry positive polarity, 39.78% indicate negative polarity, and 34.41% are neutral. A possible interpretation is that a good degree of internationalization actively boosts the candidate’s evaluation, while lack of internationalization actively hinders it, since this aspect also tends to elicit an active response from the panel. Conference participation is more often mentioned neutrally (52.72%), but when polarity is shown, it is more often positive (34.07%) than negative (13.19%). The remaining categories not only have few occurrences to make a meaningful interpretation, but they are also largely neutral in nature. Nevertheless, they are shown in the table above for the sake of completeness. Overall, it is shown that some criteria do have a degree of polarization in the sense that they are commented through polarity-loaded sentences, rather than neutral ones, suggesting that

[2] Although it was considered to split the valence analysis by panel, it was opted to conduct the analysis using the global sample for ease of presentation.

these criteria are likely to have an implicitly higher weight in the final assessment. Figure 1 summarizes the global results of this exercise:

<INSERT FIGURE 1 HERE>

**Figure 1**. Polarity and content analysis of the global data. On the left: percentage of occurrences per theme and per polarity type. On the right: frequency of occurrence of each theme.

*Analysis 2 – Scientometric analysis*

For organization purposes of this section, note that the dependent variable is always the candidate score; the various columns in the tables refer to differing counting methods for the independent variables, some of which are not available for all variables. Model I consists of raw counting on all variables; Model II only counts international material for activities and production, with all others being raw counts; Model III only counts first-authored materials for production, with all others being raw counts; Model IV only counts single-authored materials for production, with all others being raw counts; and Model V employs fractional counting for production, with all others being raw counts.

We begin the analysis by observing the results for the 3rd edition panel, which are shown in Table 5:

<INSERT TABLE 5 HERE>

The first surprising finding is that, while employing only raw counting, none of the scientometric indicators have any significant impact on the candidate score. The only significant effect is the number of unique countries where the candidate has worked ($B = 0.349$, $p < 0.05$), which correspondents to part of the internationalization criteria. However, when we move on to Model II – in which activities and production are only counted if they are of an international nature – still no significant effects emerge. This suggests that physical internationalization, as represented by international career mobility, is more important than international projection by means of publication in international outlets or dissemination of work abroad. Unique Job Country count maintains its effect. Proceeding into Model III, in which production is only counted if the candidate is the first author, still no effects emerge, suggesting that first-authoring academic materials, by itself, is also not a substantial factor affecting the candidate score. However, when we move to Model IV, in which production is only counted if the candidate was the sole author, significant findings emerge. First, single-authored book chapters significantly decrease the candidate score ($B = -0.345$, $p < 0.05$), a surprising finding, especially since these were mentioned in 37.5% of the qualitative evaluations and book chapters are a typical scientific output in the social sciences. More notably, journal articles, when only single-authored papers are counted, now have a positive impact on the candidate score ($B = 0.155$, $p < 0.01$). Additionally, both peer review activity counts ($B = 0.151$, $p < 0.05$), teaching activities ($B = 0.045$, $p < 0.05$), and projects ($B = 0.096$, $p < 0.05$) now have significant effects on the candidate score. Additionally, being a female also has a positive impact on the candidate score when only single-authored articles are counted ($B = 0.603$, $p < 0.05$), while time since the PhD has a negative effect ($B = -0.330$, $p < 0.01$). The fact that these variables were not significant with other types of production counting is interesting, especially since their own counting remained unchanged. One possible interpretation is the occurrence of an halo effect (Nisbett & Wilson, 1977) – a candidate which has a multitude of single-authored papers is likely to result in a positive appraisal in other not necessarily related aspects of the curriculum, and gender seems to also play a role – the dynamics of which are complex (e.g., Abramo et al., 2009, 2018) and out of the scope of this paper. Time since PhD is easily explained – single-authored production might be impressive at a very early career stage, but if a longer time period has passed since the PhD's conclusion, the competitive advantage is likely lost. Finally, Model V employs fractional counting of the candidate's production. In this model, Gender, Teaching activities, and Projects are no longer significant; everything else remains identical to Model IV.

Based on the five models, our interpretation is that the most consistent criterion for evaluation is the number of jobs the candidate has held abroad, reflecting the internationalization aspect based on geographical physical mobility, not scope of international activities. Production and activities, by themselves, do not meaningfully affect the candidate evaluation. However, the scenario changes when single-authored or fractional counting is employed; single-authored papers, or papers with few co-authors, have a positive impact on the evaluation with spillover effects into other aspects of the curriculum.

We proceed by observing the results for the 4th edition panel, which are shown in Table 6:

<INSERT TABLE 6 HERE>

The results for this panel are substantially different from the previous panel, in that none of the scientometric indicators have any significant impact on the candidate score, regardless of the type of counting employed. The only aspect with a significant impact is the number of countries where the candidate has worked ($p < 0.05$), which resonates with the internationalization criteria also found in the previous panel. The only other noteworthy finding is that other jobs – non-academic in nature – have a negative impact on the candidate score ($B = -0.249$, $p < 0.05$), which by itself is unsurprising, but only when single-authored counting is employed. A possible explanation would be that holding non-academic jobs would detract from publishing single-authored materials, but since the productivity indicators by themselves are not significant, the dynamics at play here are unclear at this time. Overall, despite the implicit criteria stated in the qualitative evaluations (and the explicit evaluation guidelines), there is no statistical evidence for the scientometric indicators having an actual impact on the candidate evaluations for this panel. The fact that the results differ across panels hints at the existence of "written rules" which go beyond the simple following of the explicit and implicit criteria, which can vary by panel; something which is aligned with the existing literature (Lamont, 2009).

*Analysis 3 – Mixed-methods analysis*

In this section, the results from the sentiment analysis are cross-matched with the scientometric indicators in order to ascertain whether there is a correspondence between the qualitative evaluations and quantitative productivity indicators. This is done by comparing individuals with positive, neutral, and negative polarity on various themes, with regards to the corresponding metric – for example, if individuals with a positive polarity on the "Publications" theme do in fact have more publications than individuals with neutral or negative polarity. Since not all themes emerged with equal frequency across the qualitative evaluations, and some of them are too uncommon to make a meaningful statistical interpretation, we only consider themes which have a minimum of ten occurrences in positive, neutral, and negative polarity. Thus, comparison was done only for the Publications (N = 142), Internationalization (N = 93), Conferences (N = 91), and Teaching (N = 69) themes. Publication polarity was compared using the “Production – Journal Articles” variable; Internationalization using the “Unique job countries Count”, and the sum of all international-counted activities, and also the sum of all international-counted production; Conferences was compared through the “Activities – Conference Participation” and “Production – Conference Materials” variables; and finally, Teaching was compared using the “Activities – Teaching” and “Teaching Jobs Count” variables.

Beginning with the Publications polarity comparison, significant differences were found across polarity types (F(2, 123) = 27.877, $p < 0.001$). Post-hoc testing revealed that candidates with a positive polarity on the Publications theme had a higher number of publications (M = 9.93, SD = 7.259) than those with neutral (M = 4.49, SD = 2.769) or negative polarity (M = 2.75, SD = 2.629); however, candidates with negative or neutral polarity on publications did not exhibit significant differences between themselves, as shown in Figure 2:

<INSERT FIGURE 2 HERE>

**Figure 2.** Comparison of journal article counts across publication polarity types.

Regarding the Internationalization polarity comparison, significant differences were found across polarity types for Unique job countries count (F(2, 78) = 4.498, $p < 0.05$), total international activities (F(2, 78) = 3.420, $p < 0.05$), and also total international production (F(2, 78) = 8.686, $p < 0.001$). In terms of unique job countries count, significant ($p < 0.05$) differences were found between candidates with positive polarity on internationalization (M = 3.17, SD = 1.543) and those with neutral (M = 2.10, SD = 1.319) or negative (M = 2.21, SD = 1.038) polarity. No significant differences were found between candidates with neutral or negative polarity. Concerning the total of international activities, the post-hoc test found no significant differences between candidates with positive (M = 26.61, SD = 27.364), neutral (M = 25.14, SD = 26.680) or negative (M = 13.02, SD = 10.647) polarity. This is in spite of the omnibus test identifying significant differences. Thus, the results for this variable can be deemed inconclusive. Finally, regarding total international production, post-hoc testing revealed that candidates with positive polarity exhibited significantly ($p < 0.05$) higher levels of international production (M = 19.50, SD = 17.054) than those with neutral (M = 10.10, SD = 9.961) or negative (M = 6.61, SD = 5.841) polarity. Neutral and negative polarity candidates did not significantly differ across themselves. Figure 3 summarizes these comparisons:

<INSERT FIGURE 3 HERE>

**Figure 3.** Comparison of unique job countries count, total international activities, and total international production across international polarity types.

Proceeding into the Conferences theme, significant differences were found across polarity types for both Activities – Conference Participation (F(2, 79) = 8.672, $p < 0.001$) and Production – Conference

Materials (F2, 79) = 6.131, $p < 0.05$). Post-hoc testing reveals that, for Activities – Conference Participation, candidates with a positive polarity on this category exhibited significantly ($p < 0.01$) higher counts of conference participation (M = 28.70, SD = 18.769) than those with neutral (M = 16.00, SD = 12.314) or negative polarity (M = 10.25, SD = 14.492). The latter two did not exhibit significant differences between themselves. Finally, for Production – Conference Materials, candidates with positive polarity exhibited a significantly ($p < 0.05$) higher count of conference-related production (M = 8.17, SD = 12.132) than those with neutral polarity (M = 1.95, SD = 2.679), but not when compared to those with negative polarity (M = 2.17, SD = 3.243). No differences were found between candidates with neutral and negative polarity. Figure 4 summarizes these comparisons:

<INSERT FIGURE 4 HERE>

**Figure 4.** Comparison of conference participation and conference materials across conference polarity types.

The final comparison is regarding the Teaching category. Significant differences were found across polarity types for both Activities – Teaching (F(2, 60) = 9.968, $p < 0.001$) and Teaching Jobs Count (F(2, 60) = 7.296, $p < 0.05$). For Activities – Teaching, post-hoc testing revealed that candidates with a positive polarity on the teaching category had a significantly ($p < 0.01$) higher intensity of teaching-related activities (M = 19.55, SD = 16.501) than candidates with a neutral (M = 7.07, SD = 8.086) or negative (M = 2.20, SD = 3.120) polarity, whereas the latter two did not differ significantly among themselves. As for Teaching Jobs Count, candidates with a negative polarity had significantly ($p < 0.05$) less teaching experience (M = 0.40, SD = 0.966) than those with neutral polarity (M = 1.71, SD = 1.566), but did not differ from those with positive polarity (M = 1.73, SD = 1.421). No other significant differences were found. This is summarized in Figure 5:

<INSERT FIGURE 5 HERE>

**Figure 5.** Comparison of teaching activities and teaching jobs across teaching polarity types.

A final exercise for this analysis was computing a global polarity rating for each evaluation. This was done by coding negative polarity occurrences as -1, neutral as 0, and positive as 1, and then simply summing all occurrences for each evaluation. This resulted in a “Polarity Rating” representing a global sentiment assessment of the qualitative evaluation. The resulting indicator revealed a relatively normal distribution, as shown in Figure 6. Finally, the candidate score was regressed on polarity rating, and it was found that the qualitative evaluation’s polarity rating was a positive predictor of the candidate score (B = 0.297, $p < 0.001$; $R^2 = 0.439$), which is also shown in Figure 6:

<INSERT FIGURE 6 HERE>

**Figure 6.** Left: Histogram of the Polarity Rating, with mean shown as a dotted line. Right: scatterplot of polarity rating and candidate score, with regression line.

Overall, this analysis highlights three important aspects. First, there is only a partial correspondence between the qualitative evaluations and the quantitative scientometric indicators. Second, the global polarity of the evaluation is positively associated with candidate score. And third, based on the pair-wise comparisons, there does not appear to be many differences at a scientometric level between candidates with negative or neutral polarity assessments; only those with positive polarity tend to exhibit significantly higher metrics when compared with the other two, in most but not all of the comparisons

**Discussion**

In this study, we aimed to explore the dynamics of a social sciences panel evaluation process by analyzing data from two editions of a competitive call in Portugal. Several key findings have emerged. First, the implicit criteria for evaluation – identified through the qualitative evaluations – are largely in line with the explicit criteria in the evaluation guidelines. This indicates that the panelists largely tend to follow the stated criteria to the best of their ability, even though there are some occurrences which emerged which were not explicitly stated in the guidelines; this can be simply a byproduct of the subjective interpretation which is inherent to these processes (Zhu et al., 2022). However, when we observe the scientometric data, there is little to no evidence of the candidate's scientific curriculum having an actual impact on the final evaluation, in spite of these criteria being mentioned in the qualitative evaluations. The sole exception to this is internationalization, assessed as geographical physical mobility, which is systematically significant across models. We then compared qualitative evaluations with positive versus those with negative or neutral polarity. Positive-polarity evaluations, in most cases, have more of whichever scientometric indicator was noted in the positive-loaded evaluation, when compared to negative/neutral evaluations; the exceptions to this are in the comparisons for Teaching – Job Count and International – International Activities, which show no differences across polarity types in terms of the correspondent metric. The complexity of the process becomes clearer when we consider that the overall sentiment of the evaluation does in fact influence the candidate's evaluation, in spite of the scientometric indicators not directly influencing the candidate score. Thus, this hints towards the presence of some affective mechanism at work which goes beyond the explicit and implicit criteria for decision-making.

The role of affect should not come as a surprise; indeed, it has been noted in the literature that panelists follow a set of informal rules which go beyond the explicit ones (Lamont, 2009; Langfeldt, 2004). These rules tend to emerge due to the fact that "excellence" is, in this context, very difficult to define – panelists consider that they know it when they see it, but providing a formal definition for it is far more difficult; thus, "gut-feeling" and past experience can also play an important role (Lamont, 2009) – perhaps even more than the official criteria, in some cases, which can partially explain why there is such a high disagreement among reviewers especially when consensus is not the goal (Pier et al., 2018). Indeed, even at the scientometric level, our results show differing effects across the two panels. Because of this, our interpretation is that the scientometric indicators, which should be used to judge the application, do not directly influence the score *per se*; rather, they contribute to a holistic evaluation of the candidate, which can elicit various types of affective responses (which can be indirectly garnered through the sentiment analysis of the qualitative evaluation). It is then this affective response, based on a mental representation of the candidate, which will translate into the final quantitative evaluation. In particular, this can explain why there is a mismatch between the polarity of the evaluation and the scientometric indicators in some of the metrics under study; a candidate who elicited a globally positive response can be perceived as having more of a thing than a candidate with a neutral or negative response, even when this is not the case (as occurred with Teaching Job Count and International Activities, noted above).

A possible hint towards these dynamics lies in the fact that when production is counted only when the candidate is the sole author (our Model IV in the quantitative analysis), several significant effects emerge which were not found in other types of counting, especially in variables that are still raw counted in the same model. As mentioned in the respective section, we interpret this as the halo effect (Nisbett & Wilson, 1977) – a known cognitive bias which occurs when a positive appraisal of an individual's characteristic leads to a positive appraisal of that individual's other characteristics. The halo effect has been known to occur in such evaluations due to reputation and prestige (Merton, 1996); however, in this case the candidates are all early in their career so it is unlikely that reputation played a substantial role in their evaluation. Rather, in this context, we suspect that individuals who publish a substantial number of single-authored articles elicit a positive affective response, which causes panelists to also consider other aspects of the candidate in a positive light – such as peer review and teaching – which would not be otherwise considered, even though they are stated in the criteria. As such, this suggests that each individual criterion is not considered in a vacuum, but rather as a piece of a puzzle adding up to a global representation of the individual, which is then translated into a quantitative assessment. Unfortunately, this is largely speculative based on current data, and confirming this interpretation would require additional data – notably, at the panelist level - which we have no access to.

## Conclusion

The findings of this study provide further hints towards the complex mechanisms underlying panel evaluations. The fact that scientometric indicators by themselves do not appear to have much importance in the candidate score, in contrast with the sentiment of the evaluation – which does have an impact - highlights the role of affect and informal rules as a critical part of the process. Although this has been noted previously in the literature – notably, by Lamont (2009) – this study further contributes to the topic by using a mixed-methods approach, which allowed us to identify the more subtle aspects of these dynamics which are difficult to identify through quantitative or qualitative methods alone. Additionally, as previously noted, the call under study places a much greater emphasis on the candidate's curriculum than on his or her proposal; although the literature typically concerns itself with the academic judgement of proposals, this study provides a novel contribution to the literature by showing that this judgement - when applied to the candidates themselves - seems to follow similar mechanisms (and might in fact be entwined, but this would be a topic for another study entirely).

This study has several implications. For funding agencies which employ panels of experts for review processes, it provides insights into how these panels operate. Additionally, the various techniques employed work as a proof-of-concept of how these methodologies can be used to judge the scientometric validity of panel reviews. For the candidates, it provides information which can be used to better their own applications, potentially increasing their chances of success. With that said, it should be noted that this study also has several the limitations, the first of which is its case study format. The data is limited to two panels in the social sciences for a very specific national call, and as such it is uncertain whether these findings would generalize to other calls – indeed, the results differed even across both of the studied panels. Unfortunately, collecting data with such depth as the one used in this study is a daunting task, and as such it would be very difficult to conduct a large-scale study without losing granularity of information. Because of this, it was opted to frame this exercise as a case study, which can still contribute to the field through its depth of information. Another limitation is the small sample size, which unfortunately, is simply due to the nature of the population. Even though we collected data for all candidates with public profiles, the sample size is limited by the number of applications, which by itself is not very large on a per-panel basis. We considered collecting data for other panels, but this would create an additional issue; other panels have likely different dynamics, which would require a very extensive cohort analysis since it would not be sensible to simply merge panels from differing fields of science. Despite this, this exercise should provide invaluable information for all academics who deal with the peer-review process in competitive grants, be they reviewers, or reviewees.

**Competing interests**

The author declares that he applied on both calls which are presented in this case study and was initially unsuccessful in both instances; however, following an appeal and a formal complaint, funding was granted to the author. However, the author does not consider there to be competing interests, since these calls have already closed and their results are final at present time; options for appeal are likewise also closed. As such, the publishing of these findings will have no bearing on the author's own applications for these calls.

**References**

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics*, *79*(3), 517–539. https://doi.org/10.1007/s11192-007-2046-8

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2018). The effects of gender, age and academic rank on research diversification. *Scientometrics*, *114*(2), 373–387. https://doi.org/10.1007/s11192-017-2529-1

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. *Science*, *321*(5885), 15–15.

Allison, P. D., Long, J. S., & Krauze, T. K. (1982). Cumulative advantage and inequality in science. *American Sociological Review*, 615–625.

Allison, P. D., & Stewart, J. A. (1974). Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review*, *39*(4), 596–606.

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Backes-Gellner, U., & Schlinghoff, A. (2010). Career incentives and "publish or perish" in German and US universities. *European Education*, *42*(3), 26–52.

Bourdieu, P. (1988). *Homo academicus*. Stanford University Press.

Bourdieu, P. (1999). The specificity of the scientific field. *The Science Studies Reader. Ed. Biagioli M. New York: Routledge*, 31–50.

Bozeman, B. (1993). Peer review and evaluation of R&D impacts. In *Evaluating R&D impacts: Methods and practice* (pp. 79–98). Springer.

Drisko, J. W., & Maschi, T. (2016). *Content analysis*. Oxford University Press.

FCT. (2020). *Regulamento do Emprego Científico (REC)*. https://www.fct.pt/apoios/contratacaodoutorados/empregocientifico/docs/REC_CEEC_IND_3.pdf

FCT. (2021). *Evaluation Guide—Stimulus of Scientific Employment, Individual Support Call (CEEC Ind) 4th edition*. https://www.fct.pt/apoios/contratacaodoutorados/empregocientifico/docs/CEECIND_4th_Evaluation_Guide.pdf

Frandsen, T. F., Jacobsen, R. H., & Ousager, J. (2020). Gender gaps in scientific performance: A longitudinal matching study of health sciences researchers. *Scientometrics*, *124*(2), 1511–1527. https://doi.org/10.1007/s11192-020-03528-z

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis*. Pearson Education Limited.

Horta, H. (2008). On improving the university research base: The Technical University of Lisbon case in perspective. *Higher Education Policy*, *21*(1), 123–146.

Huebner, J. (2005). A possible declining trend for worldwide innovation. *Technological Forecasting and Social Change*, *72*(8), 980–986.

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

Jerrim, J., & Vries, R. de. (2020). Are peer-reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications. *The Social Science Journal*, 1–19.

Kwiek, M., & Roszka, W. (2021). Gender disparities in international research collaboration: A study of 25,000 university professors. *Journal of Economic Surveys*, *35*(5), 1344–1380.

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Langfeldt, L. (2004). Expert panels evaluating research: Decision-making and sources of bias. *Research Evaluation*, *13*(1), 51–62.

Larivière, V., Macaluso, B., Archambault, É., & Gingras, Y. (2010). Which scientific elites? On the concentration of research funds, publications and citations. *Research Evaluation*, *19*(1), 45–53.

Laudel, G. (2006). The art of getting funded: How scientists adapt to their funding conditions. *Science and Public Policy*, *33*(7), 489–504.

Lempiäinen, K. (2015). Precariousness in Academia: Prospects for University Employment. In D. della Porta, S. Hänninen, M. Siisiäinen, & T. Silvasti (Eds.), *The New Social Division: Making and Unmaking Precariousness* (pp. 123–138). Palgrave Macmillan UK. https://doi.org/10.1057/9781137509352_7

Martin, B. R. (2011). The Research Excellence Framework and the ‘impact agenda’: Are we creating a Frankenstein monster? *Research Evaluation*, *20*(3), 247–254.

McGrail, M. R., Rickard, C. M., & Jones, R. (2006). Publish or perish: A systematic review of interventions to increase academic publication rates. *Higher Education Research & Development*, *25*(1), 19–35.

Merton, R. K. (1996). *On social structure and science*. University of Chicago Press.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201–237). Elsevier.

Mongeon, P., Brodeur, C., Beaudry, C., & Larivière, V. (2016). Concentration of research funding leads to decreasing marginal returns. *Research Evaluation*, *25*(4), 396–404.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Mueller, P. S., Murali, N. S., Cha, S. S., Erwin, P. J., & Ghosh, A. K. (2006). The association between impact factors and language of general internal medicine journals. *Swiss Medical Weekly*, *136*(2728).

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250.

Patricio, M. T. (2010). Science Policy and the Internationalisation of Research in Portugal. *Journal of Studies in International Education*, *14*(2), 161–182. https://doi.org/10.1177/1028315309337932

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 2952–2957.

Pinto, S., & Sá, M. H. A. e. (2020). Scientific research and languages in Portuguese Higher Education Institutions. *Language Problems and Language Planning*, *44*(1), 20–44. https://doi.org/10.1075/lplp.00054.pin

Roumbanis, L. (2021). Disagreement and Agonistic Chance in Peer Review. *Science, Technology, & Human Values*, 01622439211026016. https://doi.org/10.1177/01622439211026016

Sano, H. (2002). The world's lingua franca of science. *English Today*, *18*(4), 45–49.

Santos, J. M., Horta, H., & Amâncio, L. (2020). Research agendas of female and male academics: A new perspective on gender disparities in academia. *Gender and Education*, 1–19. https://doi.org/10.1080/09540253.2020.1792844

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497. https://doi.org/10.1136/bmj.314.7079.497

Shaikh, M. A. M., Prendinger, H., & Mitsuru, I. (2007). *Assessing sentiment of text by semantic dependency and contextual valence analysis*. 191–202.

Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(7), 190194. https://doi.org/10.1098/rsos.190194

Stephan, P. (2012). Research efficiency: Perverse incentives. *Nature*, *484*(7392), 29–31.

Tukey, J. (1953). Multiple comparisons. *Journal of the American Statistical Association*, *48*(263), 624–625.

Young, M. (2015). Competitive funding, citation regimes, and the diminishment of breakthrough research. *Higher Education*, *69*(3), 421–434.

Zhu, W., Li, S., Ku, Q., & Zhang, C. (2020). Evaluation information fusion of scientific research project based on evidential reasoning approach under two-dimensional frames of discernment. *IEEE Access*, *8*, 8087–8100.

Zhu, W., Li, S., Zhang, H., Zhang, T., & Li, Z. (2022). Evaluation of scientific research projects on the basis of evidential reasoning approach under the perspective of expert reliability. *Scientometrics*, *127*(1), 275–298. https://doi.org/10.1007/s11192-021-04201-9