RESEARCH ARTICLE

Gender bias in funding evaluation: A randomized experiment

Laura Cruz-Castro and Luis Sanz-Menéndez

Institute of Public Goods and Policies (IPP), Spanish National Research Center (CSIC), Madrid, Spain

Keywords: experiments, funding agencies, gender bias, peer review, research evaluation, research funding

ABSTRACT

Gender differences in research funding exist, but bias evidence is elusive and findings are contradictory. Bias has multiple dimensions, but in evaluation processes, bias would be the outcome of the reviewers' assessment. Evidence in observational approaches is often based either on outcome distributions or on modeling bias as the residual. Causal claims are usually mixed with simple statistical associations. In this paper we use an experimental design to measure the effects of a cause: the effect of the gender of the principal investigator (PI) on the score of a research funding application (treatment). We embedded a hypothetical research application description in a field experiment. The subjects were the reviewers selected by a funding agency, and the experiment was implemented simultaneously with the funding call's peer review assessment. We manipulated the application item that described the gender of the PI, with two designations: female PI and male PI. Treatment was randomly allocated with block assignment, and the response rate was 100% of the population, avoiding problems of biased estimates in pooled data. Contrary to some research, we find no evidence that male or female PIs received significantly different scores, nor any evidence of same-gender preferences of reviewers regarding the applicants' gender.

1. INTRODUCTION: GENDER DIFFERENCES AND GENDER BIAS IN RESEARCH FUNDING

There is evidence that gender differences continue to exist in many dimensions of science activities and research outcomes in all countries (Ceci, Ginther et al., 2014). Women show lower application levels, success rates, and levels of research funding (Pohlhaus, Jiang et al., 2011; Suarez, Fiorentin, & Pereira, 2023); differences across genders exist in the proportion of female researchers in many STEM careers (Bello & Galindo-Rueda, 2020); the percentage of women at the top of careers as full professors or in highly prestigious universities is lower (Directorate-General for Research and Innovation (European Commission), 2021); and women have slower career advancement (Wang & Degol, 2017), get less prestigious special chairs (Treviño, Gomez-Mejia et al., 2018), show weaker publication and citation patterns (Mayer & Rathmann, 2018), tend to collaborate less (Aksnes, Piro, & Rørstad, 2019; Gaughan & Bozeman, 2016; Kwiek & Roszka, 2021), and get higher rejection rates (Fox & Paine, 2019). In some of these domains, though, there seems to be a trend of diminishing gaps (Cruz-Castro, Ginther, & Sanz-Menéndez, 2023). Additionally, every single factor interacts with the others; for example, publications could account for differences in funding (Ginther, Basner et al., 2018) or differences in funding could predict career advancement patterns (Bloch, Graversen, & Pedersen, 2014). Moreover, a lack of clear concepts, theories, and causal models are among the factors accounting for the contradictory findings.

an open access journal

Check for updates

Citation: Cruz-Castro, L., & Sanz-Menéndez, L. (2023). Gender bias in funding evaluation: A randomized experiment. *Quantitative Science Studies*, *4*(3), 594–621. https://doi.org/10.1162/qss_a_00263

DOI: https://doi.org/10.1162/qss_a_00263

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00263

Supporting Information: https://doi.org/10.1162/qss_a_00263

Received: 25 November 2022
Accepted: 1 July 2023

Corresponding Author:
Luis Sanz-Menéndez
luis.sanz@csic.es

Handling Editor:
Vincent Larivière

The MIT Press

Gender bias as a general concept has been used in a quite broad sense, often confused with the existence of gender differences. Some research argues that there is gender bias in research funding; others claim that gender differences exist, but bias evidence is elusive and contradictory. Differences in conclusions can be explained partly because research evidence comes from different countries, timing, disciplines, context data, and methodological approaches.

Evaluation bias has been generally defined as a deviation from assessing an object based on merit (Lee, Sugimoto et al., 2013). However, in the allocation of research funding, peer review is used to evaluate the relative merits of the different applications (Hug, 2022), and the funding agency usually defines the assessment criteria, including their weighting, to score the quality, merit, or other dimensions to be assessed. It is reviewers who interpret and apply evaluative criteria, often without interrater agreement (Bornmann, 2015; Jerrim & Vries, 2023), and, in this context, impartiality in peer evaluations may not be possible, as definitions of evaluative criteria condition their interpretation and outcomes (Lee, 2012).

Controversies regarding the evidence itself have continued and, as recognized by *Science*, there is a general challenge with the use of peer review, both for reviewing papers in journals and for allocating funding resources (Pinholster, 2016; Sato, Gygax et al., 2021), and for interpretation of the outcomes in terms of bias (Cruz-Castro et al., 2023).

However, despite some of the evidence emerging from proportions and regressions associated with explanatory theories, causal mechanisms are not clearly described (Cruz-Castro & Sanz-Menéndez, 2020). The debate revolves around to what extent gender bias in assessment is the cause of the gender differences in funding outcomes or, in other terms, if the gender of applicants or principal investigators (PI) affects the outcome of the assessment.

In this context, if gender bias occurs in assessing applications for research funding (our focus of interest), it will be the outcome of the action and behavior of the reviewers' assessment, of course, under the specific institutional arrangements at the funding agency.

Our objectives in this paper are twofold: first, to examine the existence of bias against or in favor of a particular gender by measuring if the applicant's (PI) gender has an influence on the scoring of applications by the reviewers; and second, to study whether female and male reviewers differ in their assessments of female and male PI applicants.

The existence of contradictory evidence and claims regarding gender bias in research funding is not only the result of conceptual ambiguity and measurement differences; it is also the result of research designs and assumptions. As is well known, the search for evidence can be done from observational or experimental approaches (Rosenbaum, 2017).

In general, early observational approaches focused only on outcome distributions (e.g., success rates in applications by gender), and later, gender bias was generally understood as the difference in outcomes of interest that could not be explained by any other factor once the theoretically relevant ones (such as quality or performance) had been considered. Thus, most of the observational approaches interpret gender bias as the residual (Cruz-Castro & Sanz-Menéndez, 2020). Simple distributions of success rates and comparisons of evaluation scores across male and female applicants, without a proper causal approach, are a weak source of evidence about bias, even if they are introduced in regression analyses as control for other covariables, such as merit, or with instrumental variables to address endogeneity problems.

Additionally, much of the evidence has focused on the applicants or applications as the units of analysis (Cruz-Castro & Sanz-Menéndez, 2020). These approaches describe processes without actors and are quite unrealistic, as the subjects involved in evaluation (and potentially introducing the gender bias) are the evaluators. The focus on the applicants' attributes

and success rates misses the point of the reviewers as actors, and ignores the fact that bias is the product of human action.

Including the evaluators as an essential part in the analysis of bias is not only relevant for a better understanding of the processes causing gender bias; in fact, it has become an important element of the policy discourse and actions regarding gender balance among evaluators and within panels. This has contributed to a shift in analytical focus to include reviewers as units of analysis as well. In this regard, establishing causality is especially relevant for the design of sound policy interventions.

We acknowledge that some research has also focused on the activities of reviewers or panels and their features (gender composition, etc.), in line with the idea of identifying the causal mechanisms or at least the influential factors (Marsh, Jayasinghe, & Bond, 2011).

Underlying theories of the behavior of reviewers highlight different aspects; for example, some psychologists and sociologists have embraced gender role congruity theory (Eagly & Karau, 2002), homophily behavior (Lawrence & Shah, 2020; Murray, Siler et al., 2019), the matching hypothesis (J. R. Cole, 1979), gender similarity (Hyde, 2005), or the existence of gender stereotypes against women (Ellemers, 2018). These works are relevant to the present study because they highlight possible mechanisms.

Economists have followed different lines and assumptions, and have linked the differences in the evaluation behavior of women and men to other theories, for example related to gender differences regarding preferences (Croson & Gneezy, 2009), information availability (Tversky & Kahneman, 1974), or behavioral attitudes towards risk or competition (Niederle & Vesterlund, 2011).

Establishing the causal models and identifying the processes and mechanisms is an essential part of the explanation that is quite often missed (Cruz-Castro & Sanz-Menéndez, 2020; Traag & Waltman, 2022).

As the reviewers and panelists are actors interacting with objects (applications, etc.), in the process, if bias exists it is relevant to analyze the process of evaluation directly. In this regard, experimental approaches are well suited for measuring the effects of the causes by determining whether female and male PI-led funding applications are assessed differently. In principle, all other factors being identical, if gender differences in outcomes are found, it means that the "cause" is the gender of the applicant.

As our focus is on the use of peer review in the allocation of competitive funding and the effect of the applicant's gender, our research questions are

1. Are female and male PIs assessed differently?
2. Do female and male evaluators differ in their assessment of male and female PI applicants?

In other words, can we identify differences in the rating of competitive funding applications that could be associated exclusively with the gender of the PI or the gender of the reviewer? The identification strategy in the research design should clearly allow us to check whether a particular gender of the PI is the cause of the differences in evaluation outcomes. Experimentation that involves control by the researchers, and includes manipulation and randomization, is a valuable source of evidence, but its use for the analysis of evaluation for funding has been scarce. This paper aims to contribute to this debate on gender bias in research funding evaluation by broadening its empirical foundations.

The remainder of the paper is organized as follows: First, we review what some previous research, based on different methodological approaches (observation and experimentation), has produced in terms of evidence regarding gender bias in the context of the use of peer review for research evaluation; second, we justify the experiment and its contribution; third, we explain the methodology and research design; fourth, we present the results; and finally, we discuss the findings and present some conclusions.

2. PREVIOUS RESEARCH AND EVIDENCE

Experimental studies of gender differences in research funding are limited, probably due to the difficulties of access to the actors in the evaluation context; however, there are other topics and methodological approaches related to peer review, such as reviewing papers for journal publication or examining CVs for hiring candidates, which have been more extensively addressed with experimental approaches. The literature, however, is rather fragmented.

In this section, we examine previous research and evidence on gender bias emerging from peer review evaluation that can help to contextualize our study, contribute to identifying some analytical perspectives, and suggest some causal links. The previous research is presented according to the type of research design, highlighting its methodological approach *vis-à-vis* the findings.

2.1. Quantitative Observational Evidence

A few decades ago, research highlighted that the gender of researchers could be a factor influencing assessments of quality. Over the years, the evidence and conclusions about gender differences in funding allocation emerging from observational studies of research funding has been contradictory; at the same time, gender differences have reduced (Cruz-Castro et al., 2023).

The claim that female PIs have their grants funded less often than male PIs, and when they do have them funded the amounts are smaller for women than for men when compared with the available evidence, shows a more complex story, as there is also evidence of no systematic bias against women in peer review for funding allocation (Ceci & Williams, 2011; Kahn & Ginther, 2018) or that when experienced professional evaluators, with information about the applicant's competence, were involved, no bias was found (Ceci, 2018).

Since the classic study at the NSF by the Coles (Cole & Cole, 1979, 1981; Cole, Cole, & Simon, 1981; Cole, Rubin, & Cole, 1977, 1978), which found small differences in funding by gender but more related to rank or past performance, the controversy has continued. Evaluations of grant applications and success by gender in other countries have also yielded heterogeneous results.

Apparent lower success rates of women in funding applications have motivated studies in different countries. The highly cited work by Wennerås and Wold (1997) at the Swedish Medical Research Council (SMRC) examined the scores in different evaluation dimensions and correlated them with some indicators of "merit" (mainly bibliometric) or "social connections"; they found a discrepancy that women needed a higher performance to get funding and inferred a strong bias against them[1]. However, their data were not available, the analysis not replicated, and, therefore, the findings not confirmed (Levy & Kimura, 2009, p. 260). In

[1] Despite having small *n* (114 applications), the measures of mean differences did not include any standard test of differences in means or power analysis. Their contribution was to highlight the role of gender as a significant variable to consider in the model, after a comparison of female and male applicants of "similar" scientific productivity, and, especially, it helped to include the issue on the policy agenda.

fact, Sandström and Hällsten (2008), with a similar design and also data from the SMRC, showed that women actually fared better than men.

Jayasinghe, Marsh, and Bond (2003), with data from the Australian Research Council (ARC), found that gender differences in success were small and nonsignificant. Ley and Hamilton (2008) also found near-equal U.S. NIH funding success for men and women at all stages of their careers.

Comparative research is scarce, but in a classic metareview of research in several disciplines and countries, Bornmann, Mutz, and Daniel (2007) reported gender bias in grant funding. However, the finding was later reversed by Marsh, Bornmann et al. (2009), who showed a lack of effect of gender generalized across disciplines, countries, and funding agencies.

Van der Lee and Ellemers (2015a), focusing on reviewers' scores and outcomes in a Dutch NWO funding program, suggested gender influence in the research funding and deduced the existence of gender bias; however, their findings have been methodologically questioned (Albers, 2015; van der Lee & Ellemers, 2015b, 2015c; Volker & Steenbeek, 2015).

More recently, Severin, Martins et al. (2020) examined whether the gender of applicants and peer reviewers and other factors influenced peer review of grant applications submitted to the Swiss National Science Foundation (SNSF). Male applicants received more favorable evaluation scores than female applicants, and male reviewers awarded higher scores than female reviewers, but in multivariable analysis, differences between male and female applicants were attenuated.

Recent contributions in the field of gender disparities in research funding have advanced in the introduction of more complex analyses (looking not only at the gender of the applicant but also at the gender composition of teams; or differentiating between scoring and approval), the use of mixed methods (combining regression models with linguistic analysis of review reports), and the consideration of intersections between gender and research content.

For instance, Bianchini, Llerena et al. (2022), examining the reviews of a pan-European funding scheme (EUROCORES) over more than a decade, linked the outcome of the grant proposal peer review with the gender representation in research consortia; they found a gender effect in the evaluation outcomes of both panel members and reviewers, as applications from consortia with a higher share of female scientists were less successful in panel selection and received lower scores from external reviewers. Interestingly, they also analyzed the evaluative language of written review reports and although apparently reviewers did not perceive female scientists as being less competent, this was not reflected in the scoring, which was lower for consortia with higher female rates.

Relatedly, some studies with Dutch data have pointed out that lower scores do not automatically mean lower success rates. This is the finding of Mom and van den Besselaar (2022) and van den Besselaar and Mom (2022) with data from the European Research Council (ERC) starting grants and NWO, who report that women get systematically lower scores, but that this does not lead to overall bias in the outcomes (success rates); these findings were in line with Bol, de Vaan, and van de Rijt (2022).

In an interesting mixed-method approach, Larregue and Nielsen (2023) analyzed funded and unfunded social science applications submitted to a research council in Western Europe, exploring how applicants' disciplinary, thematic, and methodological orientations intersect with gender to shape funding opportunities. Their descriptive analysis shows that women's proposals were underfunded, with a relative gender difference of around 20%. Then they

use computational text and mediation analysis, and find that around one-third of this disparity may be attributed to gender differences in disciplinary focus, thematic specializations, and methodologies, as it appears that there is a devaluation of qualitative methods and, more broadly, interpretive, descriptive, and exploratory approaches in proposal assessments, areas in which women appear more specialized; this is in line with previous research about specialization (Leahey, 2006, 2007).

As a summary, some of the differences in the reported results relate to the use of different concepts of gender bias and various operationalization methods, in addition to contextual (country, funding agency, type of program) or sampling effects (Cruz-Castro & Sanz-Menéndez, 2020). Typically, analyses were carried out by using single-level models and analytic techniques such as correlation, analysis of variance, tests for proportions, or multiple regression; additionally, observational research is always subject to the standard problems of unobserved heterogeneity or endogeneity and it does not always include clear causal models. Our aim, however, is not to dismiss the contribution of these approaches, which have been predominant, but rather to advocate for more pluralistic methodological perspectives by showing the value of the experimental method for broadening the empirical bases of the study of evaluation bias.

2.2. Natural Experiments

There is a class of observational research that is inappropriately called "natural experiments" (Titiunik, 2021). Natural experiments, sometimes labeled as "quasi experimental designs" (Shadish, Cook, & Campbell, 2001), claim to take advantage of embedding or contextualizing the analysis in real-life situations. The idea of experiments is associated with control of two dimensions, namely randomization and manipulation (Barrera, Gerxhäni et al., 2023); natural experiments correspond to a type of observational research with no control by the researcher, or at best only include some form of randomization (Deaton & Cartwright, 2018).

Most of this research refers to gender bias in hiring and has focused on a relevant policy topic: the effects of the (gender) composition of the evaluation panels (controlling for quality and proximity), and their changes. The classic "matching hypothesis" regarding the applicant's and reviewer's same department (Cole, 1979) provided the basis for the homophily or same-gender preference claims (Murray et al., 2019); expectations were that increasing the number of women in evaluation panels would favor female scores or success rates. However, empirical results are far from conclusive.

Whether the gender composition of recruiting committees for university professorships matters is a question that has been addressed mostly in the context of country case studies. Taking advantage of the opportunity provided by "natural experiments" where the allocation of reviewers to specific evaluating committees was random (a lottery), Bagues and colleagues, first in Spain (Zinovyeva & Bagues, 2015), and later in Italy (Bagues, Sylos-Labini, & Zinovyeva, 2017) analyzed how a larger presence of female evaluators affected committee decision-making. Their results revealed that having a larger number of women on evaluation committees did not increase either the quantity or the quality of the female candidates who qualified. This work is relevant to the present study by showing that information from individual evaluation reports revealed that female evaluators were not significantly more favorable toward female candidates.

Witteman, Hendricks et al. (2019) used a Canada Institutes of Health funding program, which was divided into two new grant programs, to "differentiate" the effect of the intrinsic quality of the proposal from the merits of the candidate on evaluation outcomes. They argue that

gender differences in evaluation (in highly competitive funding programs) were less relevant when the proposal, and not the "caliber" or the CV of the applicant, was the focus of the assessment. Albeit interesting, their analysis did not control for the quality of the PI (e.g., with a measure of past performance); missing such a variable precludes ruling out competing explanations.

In regard to mechanisms, psychology research has pointed to the activation of stereotypes (Fiske, Cuddy et al., 2002). For instance, gender-role congruity theory (Eagly & Karau, 2002) proposes that perceived incongruity between the female gender role and leadership roles leads to two forms of prejudice: perceiving women less favorably than men as potential occupants of leadership roles and evaluating behavior that meets the prescriptions of a leader's role less favorably when it is shown by a woman.

Stereotype-based expectations may also be related to the prior segregation of the area subjected to evaluation (occupations, jobs, or fields), whereby, in male-dominated or female-dominated areas, reinforcement is expected, whereas in neutral areas, similar evaluations across genders will emerge. In this domain, Koch, D'Mello, and Sackett (2015) did a meta-analysis of research findings about workplace decisions according to the gender distribution of jobs. They found that men were preferred for male-dominated jobs (i.e., gender-role congruity bias), whereas no strong preference for either gender was found for female-dominated or integrated jobs; additionally, male evaluators exhibited greater gender-role congruity bias than did female evaluators for male-dominated jobs.

These theoretical perspectives suggest that status and statistical discrimination may operate in evaluation processes and often assume that male professors are more biased by gender-typical stereotypes than their female counterparts (Solga, Rusconi, & Netz, 2023).

The findings related to gender-role congruity and gender-homophily (McPherson, Smith-Lovin, & Cook, 2001; Murray et al., 2019) and some empirical anomalies (van den Besselaar & Mom, 2022) have prompted thinking on other potential explanations related to the impact of "social desirability" behavior (Krumpal, 2013), whereby when there are fewer women in a field, they would be favored in evaluations, and where there are more women in the area, it would be men who are favored. Probably we could also expect some effects of the gender equality policies in place in particular contexts (Stewart & Valian, 2018).

2.3. Laboratory and Survey Experiments

It is relevant to distinguish among different qualities of experimental approaches (Deaton & Cartwright, 2018). Potentially, laboratory and survey experiments make use of both randomization and manipulation. As mentioned, there is not much literature trying to measure and explain the possible existence of gender bias in the context of research funding evaluation, but there is some relevant research addressing the role of gender in the outcomes of evaluation for hiring decisions and acceptance of journal papers.

Of particular relevance for the present study is the work of Moss-Racusin, Dovidio et al. (2012) who conducted a randomized study (n = 127) to investigate experimentally whether science faculty exhibited a bias against female students in a hiring process. Science faculty from research-intensive universities in the United States rated the application materials of a student—who was randomly assigned either a male or female name—for a laboratory manager position. Faculty evaluators rated the male applicant as significantly more competent and hirable than the (identical) female applicant. The gender of the faculty participants did not affect responses, so female and male faculty were equally likely to exhibit bias against the female student.

With results to the contrary, Williams and Ceci (Ceci & Williams, 2015; Williams & Ceci, 2015) developed a series of randomized experiments on 873 tenure-track faculty (439 men, 434 women) from biology, engineering, economics, and psychology at 371 US universities/colleges, evaluating applications for tenure track assistant professorships. Applicants' profiles were systematically varied to disguise them for identically qualified women versus men. Results revealed a 2:1 preference for women by faculty of both genders across both math-intensive and nonmath-intensive fields, with the single exception being male economists, who showed no preference for one gender or another.

Experiments have expanded to other countries, and more recently Carlsson, Finseraas et al. (2021) examined the role of bias in academic recruitment by conducting a large-scale survey experiment among faculty in various disciplines from universities in Iceland, Norway, and Sweden. The faculty respondents rated CVs of hypothetical candidates—who were randomly assigned either a male or a female name—for a permanent position as an associate professor in their discipline. Their results also contradicted some previous findings (Moss-Racusin et al., 2012), because, despite the underrepresentation of women in all fields, the female candidates were viewed as both more competent and more hirable compared to their male counterparts. They concluded that biased evaluations of equally qualified candidates do not seem to be the key explanation of the persistent gender gap in academia in the Nordic region. However, the participants were the faculty in general, not those deciding or really involved in the hiring decision-making process.

Also relevant for our work, a recently published paper (Solga et al., 2023) uses a large factorial survey experiment with German university professors, and studies whether male and female committee members evaluate female and male applicants for professorships differently. They found neither differences between male and female professors nor the presence of a Matilda effect[2], but some advantage for female applicants in the invitation phase. They also considered that findings are probably related to the gender equality policy of having a substantial female quota in selection committees. The overall methodological approach is in line with ours, but their focus is on hiring not on funding.

These previous works have in common the construction of two groups of evaluators who received the same materials, and the randomization of the assignment of the gender of applicants to those evaluators, therefore building a group receiving female applications and a group receiving male applications. In this paper, we take a similar approach. The approach is different to studying gender blinding, where what is tested is the effect of concealing information about the gender of applicants versus otherwise[3].

Papers evaluated for journals or conferences have also long been the subject of experimental manipulation, but again, the results are partially contradictory, mainly as a consequence of the research designs (Lloyd, 1990; Paludi & Bauer, 1983).

For example, classic studies reported gender effects against female authors, but others show small or no gender bias; Borsuk, Aarssen et al. (2009) manipulated a published article to reflect different author designations. The article was then reviewed by referees of both genders at various stages of scientific training and experience. Name changing did not

[2] Consisting in female researchers receiving less recognition for their collaborative work than male researchers.

[3] We could not have taken the blinding approach, because the experiment was designed to be realistic and the gender of the PI, together with the gender balance of the team, were part of the formal evaluation criteria of the call studied.

influence acceptance rates or quality ratings. However, female postdoctoral researchers were the most critical referees, regardless of the author name provided. Additionally, there was no evidence of same-gender preferences. This study strongly suggests that more experienced women may apply different expectations to peer review, as others found in observational studies (Cruz-Castro & Sanz-Menéndez, 2021).

As we have mentioned, there is limited experimental evidence emerging from the gender effects in peer review for research funding, but the NIH in the United States has attracted some attention. Looking at the interaction between gender and race, Forscher, Cox et al. (2019) used 48 NIH R01 grant proposals and modified the PI names to create separate versions of each proposal (White male, White female, Black male, and Black female). They found little to no race or gender bias in initial R01 evaluations. Focusing on race, Nakamura, Mann et al. (2021) tested the specific effects of anonymization (concealing the race and identity) on the scores of Black and White applicants to the US NIH, with the aim of investigating bias against the former. Designed as a test of whether blinded review reduces racial disparities in peer review, they found, interestingly, that it changed the scores of White PIs' applications for the worse, but did not, on average, impact the scores of Black PI applications. Although statistically small, differences remained in favor of White PIs' scores, but anonymization reduced that difference.

As becomes evident from the data presented, careful examination of the experimental findings is also needed, not just because the findings usually relate to small *n* and have limited external validity, but mainly because most of the differences in findings come from the specifications of the research design and the inductive causal analysis.

One of the standard criticisms of laboratory or survey experiments is related to the nonrealistic nature of participants, a common feature in the majority of laboratory experiments; in field experiments or factorial surveys, what is often criticized is the absence of a real context related to the task. We aim to contribute to the literature by addressing some of these shortcomings.

3. JUSTIFICATION OF THE FIELD EXPERIMENT AND ITS CONTRIBUTION

The standard problems of confounding, lurking, or unobserved heterogeneity have not always been properly addressed in previous observational approaches on research funding and peer review. Research addressing the links between research funding and gender bias has often faced problems of validity, both external validity (as it usually referred to a single country, funding agency, or funding instrument) and internal validity, more related to the research design and the causality approach.

Acknowledging that experiments (control via manipulation of treatments and randomization) will not solve all problems of causality either (Deaton & Cartwright, 2018; Knight & Winship, 2013), we believe that, if properly designed, they can contribute to testing the existence of some regularities when assessing the bias of reviewers in favor of or against a particular PI gender, and to understanding the causal processes (Bendiscioli, Firpo et al., 2022).

At the very least, the influence of the gender of the PI within the peer review practice should be tested to ensure that the general assumptions about peer review objectivity are well founded. With an "experimental" approach, we can focus mainly on the measurement issue and be less dependent on theories and assumptions.

A need for better conceptualization has recently been highlighted in relation to peer review (Derrick, 2019; Hug, 2022), but behind the problems regarding the evidence we have

identified, there is also a lack of explicit models of causality (Cruz-Castro & Sanz-Menéndez, 2020; Traag & Waltman, 2022; van den Besselaar, Mom et al., 2020).

For the sake of clarity, by *gender disparity* we understand a difference in the outcome of interest between male and female applicants; whereas by *gender bias*, we understand any difference between male and female applicants that is *directly* causally affected (and directly measured) by their gender; a gender disparity may be the result of an indirect causal pathway from someone's gender to a particular outcome and may be affected by differences in merit, but a gender bias is a *direct* causal effect of the action of reviewers.

Even when concepts are clearly established and are part of rigorous analytical models, and the focus is on the "causes" of "gender bias," most research treats it as a "nonobservable" factor.

To highlight the differences in the analytical approaches between observational and experimental approaches we sketch them in stylized forms.

In Figures 1 and 2 we represent two simplified examples of the underlying causal models in the observational and experimental approaches, and the process of investigation included in the research design.

In typical observational research, the question of whether an applicant's gender explains differences in evaluation results is addressed by taking into consideration differences in merit by gender (measured in different forms). Bias, which is often implicit and nonobservable, is considered in this example as the residual that is not explained by the observable variables introduced in the model (unobserved heterogeneity).

Monitoring whether the gender of the applicant influences the outcome or the scoring usually includes a control for merit (e.g., publications), and, if differences in success or in the levels of merit needed to get the grant are found, as for example in the pioneering work of Wennerås and Wold (1997), the standard conclusion is the existence of gender bias. In Figure 2, due to a research design that equalizes merit and the randomization of the gender of applicants, the relations between the gender of the PI and the scoring as the outcome of the assessment could be seen as direct. And if differences are found between gender scoring, those are logically assumed to represent the existence of bias.

In the typical experimental approach, merit is made identical for all applicants and it does not vary by gender (this is why merit is absent in the figure). Therefore, the occurrence of bias can be inferred from differences in scoring of the randomly allocated male- and female-led applications.

To make it clearer, we will paraphrase the title of a paper (Dawid & Musio, 2022): Although observational researchers focus on the "causes of effects," that is, reconstructing how different explanatory variables contribute to a certain observed outcome (e.g., what are the causes of the differences), experimentalists are interested in the "effect of causes," that is, the causal effects of an intervention.

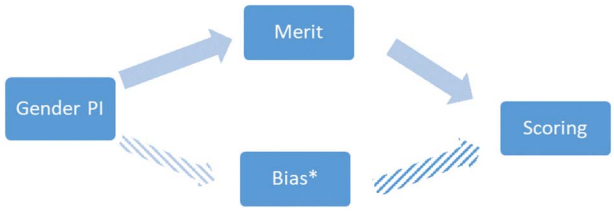

Figure 1. Observational design: Example of competing causal mechanisms of the effect of gender. *Nonobservable.

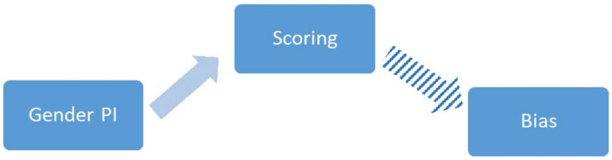

Figure 2. Experimental design: Example of causal mechanisms of the effect of gender.

In principle, the main goal and contribution of experimental approaches is to provide an explanation for previously established social regularities and to answer questions to test hypotheses from theoretical models (Barrera et al., 2023; Gërxhani & Miller, 2022). As a first rationale, the models themselves contain the causality links and determine what the experimenter needs to manipulate; a second rationale of experimental approaches is to empirically establish the existence or not of that regularity, a purpose that is of special relevance in our topic of interest.

The ideas of control and of intervention or manipulation of a variable (cause) are at the core of the experimental approach. Basically, what we do with an experiment is to modify one or more independent variables and measure the changes produced in the dependent variable of interest afterward.

Although some experimental research designs have tried to replicate peer evaluation procedures, these approaches have been criticized for their lack of realism, and experimental studies in funding agencies are rare.

To address the empirical research question of whether reviewers in funding agencies assess and score male and female applicants differently, in this work we advocate for field experiments to be embedded in research funding contexts that resemble reality as much as possible, allowing researchers to directly manipulate, allocate, observe, and test whether the gender of the applicant plays any role in the scores given to funding applications by the reviewers.

With this goal in mind, we present a field experiment implemented during the evaluation process, with the same evaluators that the evaluation agency selected for the assessment task, and with the same evaluation criteria and scores used by the funding agency. Embedded experiments, wherein theoretically relevant variables are systematically manipulated in the field, have important benefits for improving causal inference—a critical component in the development of any research field.

We seek to contribute to a literature that suffers from a high degree of inconclusiveness and contradictory evidence. We are not aware of field experiments implemented in the context of evaluation for the allocation of research funding in real time and context.

4. DESIGNING AN EXPERIMENTAL APPROACH IN A REAL FUNDING AGENCY

The randomized studies developed in the real world are usually called *field experiments*, where the word *field* comes from the original uses in agricultural research. But *field* in social research refers to "setting," and the setting or the "place" is just one relevant criterion to assess the experimentation. There are other factors that determine the "degree of fieldness" over various dimensions; some of the most important are (Gerber & Green, 2012):

1. authenticity of the treatments: whether the treatment used in the study resembles the intervention of interest in the real world;
2. the realism of participants: whether the participants in the experiment resemble the actors that usually participate in this type of process;

3. the genuineness of the context: whether the context within which subjects are receiving the treatment resembles the context of interest; and
4. the truth of the outcome measures: whether the outcome measures resemble the actual outcome of theoretical or practical interest.

We know that the strength of the experiments refers to internal validity, but this "naturalistic" approach has sometimes been presented also as a way to deal with some unforeseen threats to validity and inference, also mostly related with external validity and generalizability (Cook & Campbell, 1979; Shadish et al., 2001), which arise when drawing inferences based on the laboratory setting. Of course, generalizability is also related to other relevant factors associated with institutional and cultural context, in addition to the population or sample sizes.

From what we know, the implementation of field experiments in research-funding organizations has been limited, usually because of the complexities of dealing with ethical issues in experimentation (Hansen & Tummers, 2020), and in the interest of not affecting the fairness of the allocation processes (Rayzberg, 2019). This is probably why many of the experimental approaches on research funding have been set in somehow "unrealistic" or artificial contexts (Eden, 2017).

Our research questions are

1. Are female and male PI applications for funding assessed differently?
2. Do male and female evaluators differ in their assessment of male and female PI applicants?

To answer the questions empirically, we implemented the field experiment through a web-based factorial survey (Auspurg & Hinz, 2015) to all evaluators appointed by a funding agency (realism of participants) to assess the applications to a research funding instrument.

In our field experiment, we took advantage of the overall organization of the evaluation process to implement the experiment in the same period in which the reviewers were evaluating the real submitted applications to the call (genuineness of the context).

The experiment was embedded in the process of the evaluation of, and simultaneously with, a funding call for university research groups (Consolidated and emerging research groups of the Galician University System) of the Galician Regional Government in Spain set by the General Secretariat for Universities (SXU)[4]. The evaluation process was arranged and organized by the ACSUG (Galician Agency for the Quality—and Accreditation—of the Regional University System)[5].

We embedded the experiment in a survey to all reviewers in June 2022 (the same month in which they were doing the real evaluation work[6]). The general objective of the survey was to

[4] Information about the call can be found at https://www.edu.xunta.gal/portal/es/node/36119 (accessed April 8, 2023). For more information about the description of the funding program of the SXU, the criteria and weighting for evaluation of applications and the evaluation procedure set, see the Supplementary material (SM 1).

[5] For the Evaluation Unit (ACSUG) ethical standards, see https://www.acsug.es/ (accessed April 8, 2023).

[6] After careful consideration about the ways in which it could affect the evaluation process, we ruled out (in agreement with the funding agency) the possibility of introducing a "fictitious" additional application among the real set of applications assigned to each reviewer. Instead, we proceeded with the built-in-survey experiment. Respondents were aware that the description of the application attributes they were rating in the factorial survey was a "hypothetical one."

analyze the opinion of reviewers about the evaluation process and the appropriateness of the criteria for merit assessment defined in the call[7].

In the experiment, we asked them to score a hypothetical application[8] based on the description of some attributes relevant for the assessment, with the same definition of the evaluation criteria, using an evaluation template identical to the one used by the funding agency (truth of outcome measures). Our application consisted of three main parts: first, a description of the group composition, structure, interdisciplinarity, and gender balance, including whether leadership was female or male, with no names provided[9]; second, a quantitative summary of the group curriculum vitae (CV) from the last 3 years, including a number of past record items, such as publications (number, type and impact indicators), talent (PhD training and attraction of ERC grantees), and scientific and transfer activity (funded projects, contracts, income, patents, spinoffs); and third, a statement of the group strategy, where quality and feasibility were to be assessed[10].

It is important to emphasize that the program aim was to competitively provide research groups with basic funding; it did not fund specific research projects, and this is why the main basis of the evaluation was the CV of the group, its past records, and the group strategy statement. The scoring sheet of the experiment was the same as the one used in the program. See the weighting of the evaluation criteria in the Supplementary material (SM 1, Table S1).

The experiment design resembles some "classic" ones implemented for hiring candidates (Ceci & Williams, 2015; Moss-Racusin et al., 2012; Swim, Borgida et al., 1989; Williams & Ceci, 2015) in the sense that a single and unique application (and its merits), which is characterized by specific attributes related to the evaluation criteria, is embedded into a population survey experiment (Mutz, 2011) or factorial survey (Auspurg & Hinz, 2015).

[7] For the research design, a specific challenge is to find a balance between participants' informed consent and describing the substantive issue of interest in the survey. Methodologically, it is true that if subjects of the experiment (field survey in our case) were aware of the substantive topic (a quite sensitive one, as it is gender bias) they could change their behavior (implicit bias or stereotyped beliefs) (Krumpal, 2013) to adapt to the more socially desirable behavior (Walzenbach, 2019), or just opt for a self-selection behavior of not answering (Leeper, 2017). It is well known that not considering this in the research design is very problematic for sustaining the robustness of the conclusions. This is why we embedded our randomized experiment in a general study of the evaluation practices in academia, specifically implemented in one evaluation agency, and focused on one funding instrument design. In factorial surveys, the randomization of treatments (of some questions or parts of the questionnaire) has become standard. More information about reporting on the experiment is provided in the Supplementary material (SM 2).

[8] Herein, experimental application.

[9] There is a factor related to the funding instrument under study (research group funding) that is worth highlighting: the fact that the gender composition of the group was included as an evaluation item in the template; as mentioned before, this precluded the design of an experiment based on gender blinding, but at the same time it allows us to use this feature for designing the measurement of the effect of gender.

[10] The experimental application resembled the real ones that the reviewers were in parallel evaluating for the agency but was not completely realistic. This is for two reasons: First, as the experiment was embedded in a survey, there were space and time limitations for completion; and second, and most importantly, stratifying by field was not feasible due to the size of the population of evaluators. Therefore, for the experimental application to be generic, and to fit in a variety of potential scientific fields, most information of the group CV had to be presented as quantitative indicators, although some indicators of quality and impact about their publications were introduced in the form of publications in Q1, position of papers in distribution of citations, or number of prestigious grants, such as those from the ERC, to mention some examples. Overall, the experimental application was shorter and contained mostly quantitative information, but we do not believe that this relative lack of authenticity compromised the validity of the experiment.

There were two versions of the same identical experimental application, with a description of the merit items (past performance), strategy, and group and PI characteristics to be assessed, which varied only in the designation of the gender of the group leader or PI. To test whether reviewers assessed male and female PIs differently, a group randomly received a female-led application (treatment), whereas the other group received a male-led application (control); we use the terms *treatment* and *control* in the standard way in the experimental literature.

One of the challenges for the experimental design is to rule out the possibility of a "bad randomization" allocation that could produce treatment and control groups whose background attributes differ in some relevant way; this situation can produce errors of inference if pooled.

Considering we had a fairly small population of 74 reviewers[11] selected and an unequal distribution by gender, but inside the boundaries of what the law determines as a gender balanced evaluation panel and reviewers' composition, it was not advisable to use a completely random assignment, where each unit (reviewers) is assigned to a treatment (female applicant) and control group (male applicant) with equal probability. To favor the conditions of same probability of assignment, we used a blocked assignment (see flow diagram in Figure 3), where observations are first divided into two distinct strata (male and female reviewers) and the subjects within each stratum are allocated randomly to treatment and control groups. With this approach, male and female reviewers have the same probabilities to get either of the two designated PI genders, and each block could also be analyzed as an independent experiment[12].

In summary, block randomization guarantees that a specific proportion of a subgroup of the population will be assigned to the treatment and control groups (see Figure 3). This design feature is important, because unless the probability of assignment to the treatment group is identical for every block, pooling observations across blocks will produce biased estimates of the overall average treatment effect (ATE). Advantages also relate to the practical or ethical imperatives and to statistical precision, which is very important in small populations, as is the case here.

For the randomization, subjects were listed in a spreadsheet in two groups according to gender and assigned a random number in each group. Assignment to the two versions (designation of Female PI- or Male PI-led team) was randomized, in a blocked assignment by

[11] We need to clarify that this is not a small sample study (typically known as small N design) (Smith & Little, 2018). We are not drawing a random sample from an unknown population but instead working with the whole small population of a funding agency case. This difference has implications because our design concentrates its experimental power at the agency case level and provides high-powered tests of effects at that level. Working at the organizational and program level has advantages in terms of precise measurement, experimental control, and replication. We believe that in environments or contexts that can be explored at the individual level (as in the case of remote reviews) and when our focus (bias) is on aggregate results for our population, studies with a relatively limited number of participants are less prone to criticism regarding size.

[12] Block randomization ensures that equal numbers of male and female evaluators are assigned to each experimental condition. This design has been implemented to address some practical statistical concerns: First, the reduction of sampling variability if all is left to a complete or simple randomization procedure, considering that male and female evaluators did not represent the same proportion in the population. With block randomization, the subjects in each block (men and women) have similar potential outcomes. Second, block randomization also allows that subgroups are available for separate analysis; for example, in our gender analysis we might be interested in comparing the ATE (average treatment effect) among men with the ATE among female reviewers (Imbens & Rubin, 2015). For a snapshot description of our experiment see the Supplementary material (SM 3).

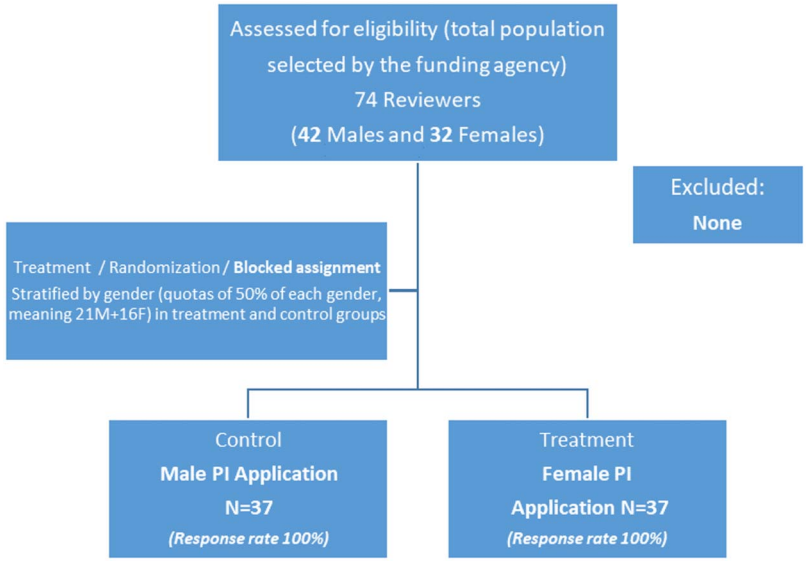

Figure 3. Flow diagram of the field experiment.

reviewer gender, using a pre-established algorithm for randomization produced by a third-party source.

We are measuring the existence of bias in the evaluation of applications of female and male PIs. Results will be examined in terms of the scores (0 to 5 in the relevant item) that the specific item of "group gender composition and leadership" has received from male and female evaluators in cases in which the PI of the application is shown as male or female. The applications being identical in all other respects, if mean scores (ATE) are significantly different between male and female designated PIs, we could empirically confirm the existence of some kind of gender bias, caused by the gender of the PI.

In sum, we believe we add to the existing literature by

1. implementing the experiment in a genuine evaluation context in a funding agency;
2. studying a real population of evaluators selected by the funding agency; and
3. using the whole population involved and not a sample.

5. MAIN RESULTS OF THE FIELD EXPERIMENT

In this section, we present the main results from the field experiment developed in the funding agency. We will focus first on the ratings by the reviewers of the evaluation item that includes designation of the gender of the PI. Second, we will analyze how reviewers from both genders assessed the applications from female and male PIs.

5.1. Are Female and Male PIs Assessed Differently?

When we analyze the evaluation item that includes identification of the gender of the PI, we find some relevant results.

As Figure 4 shows, evaluators tend to score male PIs higher, although the differences are small and not statistically significant at 95% CI. Therefore, when reviewers are randomly assigned to the treatment, there are no significant differences in the rating of the application

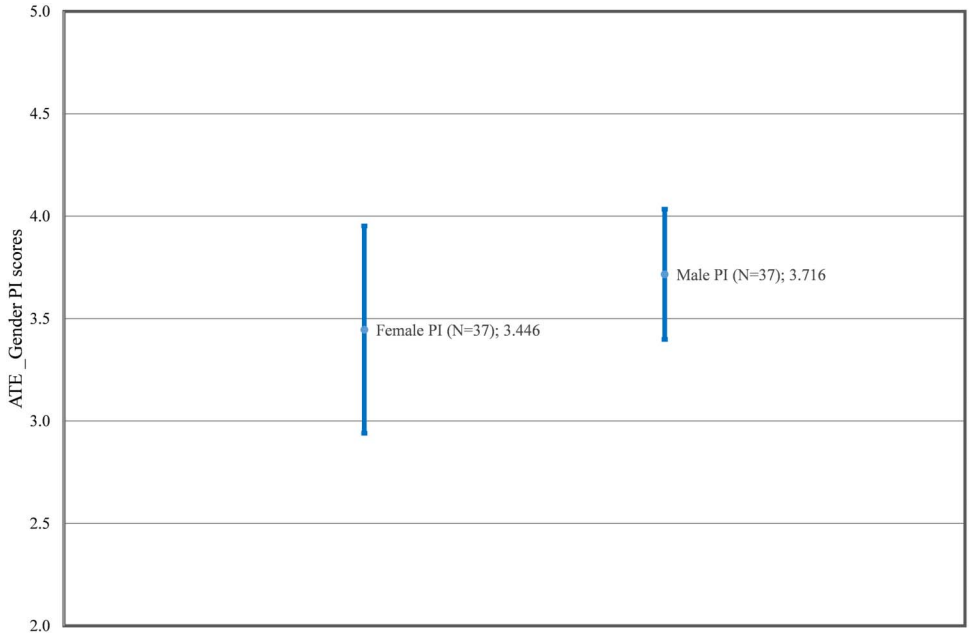

Figure 4. Mean scores (ATE) for the randomly assigned gender PI and CI at 95%.

leadership item for male and female PIs. The mean score for female PIs is 3.446 (SD 1.5174), and the mean score for male PIs is 3.716 (SD 0.9541).

This visual conclusion is also confirmed if we treat our cases as a sample of an unknown population, with various tests to check the levels of uncertainty of our measures[13] (Cox, 2020): First, with the *t*-test assuming equal variances as a parametric tool; second, with the Kruskal-Wallis test as nonparametric; and, finally, with bootstrap confidence intervals for effect size[14].

The *t*-test of the mean differences of the scores given by the evaluators to the leadership item shows no significant differences between male or female PIs (see Table 1).

The *p*-value associated with the *t*-test is not lower than 0.05—the contrast is not significant, in bilateral or left or right unilateral; this clearly means that the dependent variable (the scores taken as a continuous variable) does not present differences in the mean among the two groups, with a level of confidence of 95%. Thus, we cannot reject the null hypothesis (H_0) that no differences exist between both groups.

Whereas *p*-values are used to assess the statistical significance of a result, we also need to make an additional estimation of the effect size. Measures of effect sizes are used to assess the practical significance of a result. The effect sizes most commonly used refer to measures of group differences (the *d* family); the *d* family includes estimators such as Cohen's *d*, Hedges's *g*, and Glass's *d* (see Table 2).

For female PIs, the effect size for scores of the leadership item are 0.21 standard deviations (SD) lower than for male PIs. This result is usually considered a small effect size, as it represents around 0.2 standard deviations (Cohen, 1988); in another possible interpretation, based

[13] We are aware that some statisticians question the emphasis on approaches focusing on the rejection of the null hypothesis, statistical significance, and the confidence intervals (Cox, 2020; Fraser, 2017), but we used them as a way of addressing our research questions.

[14] For all the estimates presented in the paper we have used STATA 17 (StataCorp, 2021).

Table 1. *T*-test (two-sample *t*-tests with equal variances) of gender PI scores. Mean values, SE, SD, and CI at 95% of the scoring

Group	Obs.	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Female PI (1)	37	3.4459	0.2495	1.5174	2.9400	3.9519
Male PI (2)	37	3.7162	0.1569	0.9541	3.3981	4.0343
Combined	74	3.5811	0.1472	1.2661	3.2878	3.8744
diff		−0.2703	0.2947		−0.8577	0.3172
diff = mean(1) − mean(2)			t = −0.9172			
H_0: diff = 0			degrees of freedom = 72			
H_a: diff < 0		H_a: diff != 0		H_a: diff > 0		
Pr(T < t) = 0.1811		Pr(T > t) = 0.3621		Pr(T > t) = 0.8189		

on the *U*-statistics, it means 15% nonoverlap and 85% overlap in the distributions, or that 54% of the group of male PIs exceeds the 50th percentile of the female PI group.

As the scoring could also be taken as categorical, and the distribution is not completely normal, a nonparametric test was implemented. A Kruskal-Wallis *H* test was conducted to confirm, under different assumptions, if the rating of the leadership was different: (a) Female PI (n = 37); (b) Male PI (n = 37). The test showed that there was not a statistically significant difference in scores between the two groups, $\chi^2(1) = 0.112$, $p = 0.7375$.

A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no real difference. In theory, if the *p*-value is less than or equal to the significance level, we should reject the null hypothesis and conclude that not all population means are equal. As the *p*-value is much larger than the significance level, H_0 cannot be rejected. Therefore, the null hypothesis cannot be rejected with this test either.

Simulation studies have shown that bootstrap confidence intervals may be preferable to confidence intervals based on the noncentral *t* distribution when the variable of interest does not have a normal distribution (Algina, Keselman, & Penfield, 2006; Kelley, 2005)[15].

We also did a nonparametric bootstrap estimation resampling the cases. Results from the bootstrap statistics (see Table 3) are robust with the other effect size test implemented.

5.2. Do Female and Male Reviewers Differ in Their Assessments of Male and Female PI Applicants?

We also tested if there are differences in the patterns of evaluation by female and male evaluators. In our experiment, male and female reviewers are assigned a female or male-led application with the same probability.

As groups, male and female reviewers do not score significantly differently; in aggregate, female reviewers score slightly lower (3.48, SD 1.35) than male reviewers (3.65, SD 1.21), but the differences are small and, again, not statistically significant.

[15] As is known, if the *p*-value associated with the Shapiro-Wilk normality test statistic is greater than 0.05, the assumption that the dependent variable is distributed as normal is met. As the *p*-value associated with the homoscedasticity test statistic (homogeneity of variances) is not greater than 0.05, this initial assumption is violated, and it is not recommended to proceed only from a parametric point of view.

Table 2. Effect size based on mean comparison of scores for gender PI

Effect size	Estimate	[95% Conf. Interval]	
Cohen's *d*	−0.2132	−0.6695	0.2445
Hedges's *g*	−0.2110	−0.6625	0.2419
Glass's Delta 1	−0.1781	−0.6344	0.2806
Glass's Delta 2	−0.2833	−0.7417	0.1790
Point-Biserial *r*	−0.1075	−0.3214	0.1230

We used a nonparametric test, and the Kruskal-Wallis test showed that there was not a statistically significant difference in scores between the two groups of evaluators, $\chi^2(1) = 0.150$, $p = 0.6985$.

But the interesting question we can address thanks to the block assignment is the interaction between reviewer gender and PI gender: Do female reviewers and male reviewers score differently when assessing the female and male PIs? As we observe in Figure 5, the difference in the mean values among the groups is mainly related to a different pattern of the distribution of marks, in which the marks assigned by female reviewers to female PIs (in comparison with male PIs) show a higher dispersion than the marks assigned to male PIs (who are rarely given very low scores by female reviewers).

This is not the case for male reviewers; analyzing the mean values of the scores assigned by male reviewers to the randomly allocated applications (male or female PI), we observe similar median values and a lower dispersion of marks (see Figure 5).

As noted, the differences in the aggregate mean values of scores arise mainly from the different way in which female reviewers assess female and male PIs, especially at the extreme of the distribution.

The advantage of the stratified experimental design or blocked assignment is that we can also monitor the scoring patterns of female and male reviewers and treat them as independent experiments.

Figure 6 represents the mean scores of the treatment differentiated by the gender of the reviewer. Analyzing the mean values, we observe that female reviewers assign higher scores to male PIs (3.81, SD 0.8539) than to female PIs (3.16, SD 1.6705); in comparison, the mean rating of female and male PIs by male reviewers is almost the same (3.67 and 3.64), with an SD of 1.3904 for female PIs and an SD of 1.0385 for male PIs, respectively.

Table 3. Bootstrap estimation

					Observations: 74 Replications: 1,000	
_bs_1: r(d) _bs_2: r(g)	Observed bootstrap Coef.	Std. Err	*z*	*P* > |*z*|	Normal-based [95% Conf. Interval]	
_bs_1 |	−0.2132	0.2506	−0.85	0.395	0.7043	0.2779
_bs_2 |	−0.2110	0.2479	−0.85	0.395	0.6970	0.2750

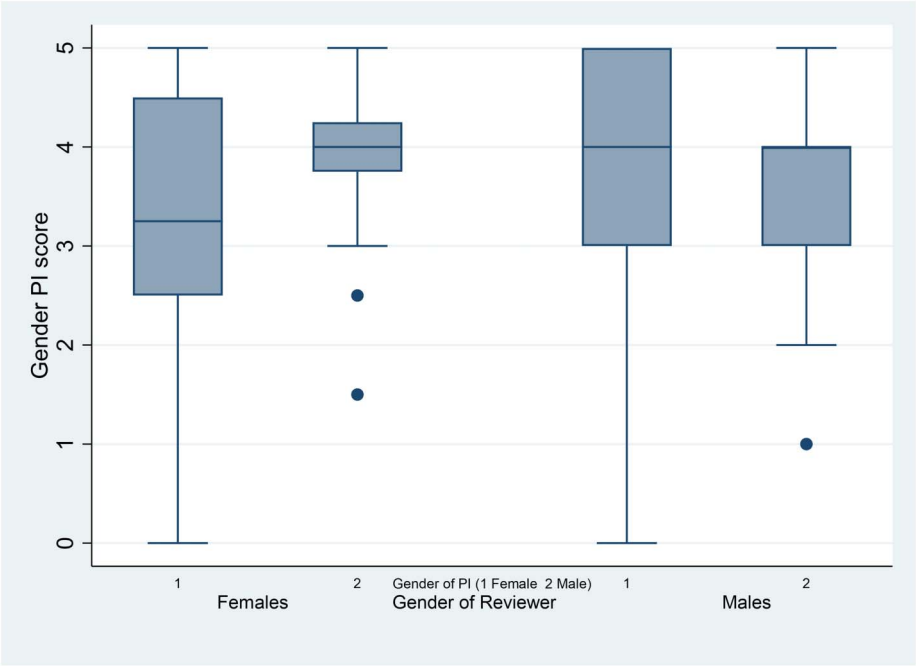

Figure 5. Box plot distribution of scores of Female PI and Male PI by gender of the reviewers. Note: From left to right: Female reviewer to Female PI (1), Female reviewer to Male PI (2), Male reviewer to Female PI (1), Male reviewer to Male PI (2).

Male reviewers rate male and female PIs with almost the same mean scores, and female reviewers rate female PIs worse than male PIs, who are favored.

As in the first analysis, we conducted some additional tests to assess the uncertainty of our measures. Again, assuming that our population was a sample, the differences in the two-

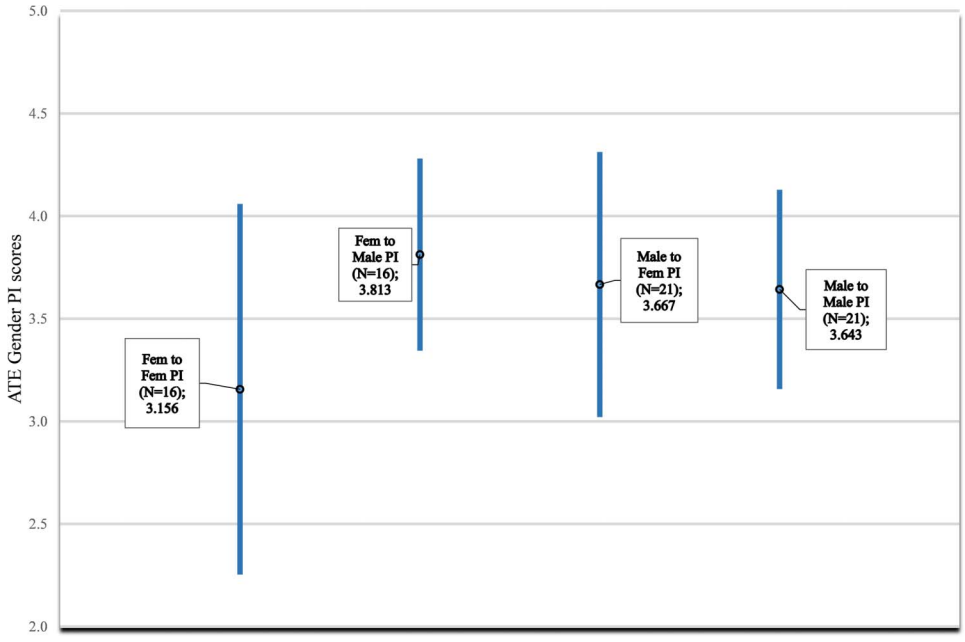

Figure 6. Mean scores (ATE) of gender PI by gender of the reviewers and CI at 95%.

Table 4. Two-sample *t*-test with equal variances of score differences of the gender of PI by gender of the reviewers

Female reviewers						
Group |	**Obs**	**Mean**	**Std. Err.**	**Std. Dev.**	**[95% Conf. Interval]**	
Female PI (1)	16	3.1562	0.4176	1.6705	2.2661	4.0464
Male PI (2)	16	3.8125	0.2135	0.8539	3.3575	4.2675
Combined	32	3.4844	0.2381	1.3469	2.9987	3.9700
diff |		−0.6562	0.4690		−1.6141	0.3016
diff = mean(1) − mean(2)			t = −1.3992			
H_0: diff = 0			degrees of freedom = 30			
H_a: diff < 0		H_a: diff != 0		H_a: diff > 0		
Pr(T < t) = 0.0860		Pr(T > t) = 0.1720		Pr(T > t) = 0.9140		
Male reviewers						
Group |	**Obs**	**Mean**	**Std. Err.**	**Std. Dev.**	**[95% Conf. Interval]**	
Female PI (1)	21	3.6667	0.3034	1.3904	3.0337	4.2996
Male PI (2)	21	3.6429	0.2266	1.0385	3.1701	4.1156
Combined	42	3.6548	0.1870	1.2122	3.2770	4.0325
diff		0.0238	0.3787		−0.7416	0.7892
diff = mean(1) − mean(2)			t = 0.0629			
H_0: diff = 0			degrees of freedom = 40			
H_a: diff < 0		H_a: diff != 0		H_a: diff > 0		
Pr(T < t) = 0.5249		Pr(T > t) = 0.9502		Pr(T > t) = 0.4751		

sample *t*-test with equal variances are not statistically significant, but there are some important nuances (see Table 4).

For both female reviewers and male reviewers taken as groups, as the *p*-value associated with the *t*-test is not lower than 0.05, the contrast is not statistically significant, whether it is bilateral or right or left unilateral, with a confidence interval of 95%. As a conclusion, the dependent variable does not present differences in means between the two groups tested (female and male reviewers scoring female and male PIs), with a confidence interval of 95%.

As in the analysis of the scores of the PI leadership item, we will repeat the nonparametric analysis, considering the distribution of the ratings, for the four different groups of interest (evaluators' gender and PIs' gender).

A Kruskal-Wallis *H* test was conducted to determine if the rating of the leadership item was different for the four groups: (a) Female Reviewers Assessing Female PIs (n = 16); (b) Female Reviewers Assessing Male PIs (n = 16); (c) Male Reviewers Assessing Female PIs (n = 21); and (d) Male Reviewers Assessing Male PIs (n = 21). The Kruskal-Wallis test showed that there was not a statistically significant difference in scores between the four groups, $\chi^2(3) = 1.398$,

Table 5. Effect size based on mean comparison

Score of female reviewers effect size of mean comparison of PI gender			
Effect size	Estimate	[95% Conf. Interval]	
Cohen's *d*	−0.4947	−1.1948	0.2134
Hedges's *g*	−0.4822	−1.1646	0.2080
Glass's Delta 1	−0.3928	−1.0934	0.3202
Glass's Delta 2	−0.7685	−1.5022	−0.0129
Point-Biserial *r*	−0.2475	−0.5251	0.1095
Score of male reviewers effect size of mean comparison of PI gender			
Effect size	Estimate	[95% Conf. Interval]	
Cohen's *d*	0.0194	−0.5856	0.6242
Hedges's *g*	0.0190	−0.5745	0.6124
Glass's Delta 1	0.0171	−0.5880	0.6218
Glass's Delta 2	0.0229	−0.5823	0.6275
Point-Biserial *r*	0.0099	−0.2874	0.3046

p = 0.7059. Therefore, we cannot reject the null hypothesis (H_0) that no differences exist between the evaluation of female and male PIs by female and male reviewers.

Although we could not reject the null hypothesis, we observe more differences in the effect size based on the mean comparison of the scoring of female reviewers on female and male PIs than in the effect size of the male reviewers doing the same (see Table 5).

Comparing the effect size (Cohen's *d*) of the gender of the PI for female and male reviewers, we observe important differences in scoring by female and male reviewers. Although the effect size of the differences in scores of the female reviewers when assessing female and male PIs goes up to almost 0.50 standard deviations, the effect size of the differences in scores of the male reviewers when assessing female and male PIs is almost marginal: 0.02 standard deviations.

An effect size of 0.5 means that the score allocated by female reviewers to male PIs is 0.5 standard deviations higher than the average female PI and, hence, male PI scores exceed the scores of 49% of the female PIs.

With these analyses, we can also confirm that the source of the small difference identified in the ratings across the gender of the PI scoring has its origins in the differences in the scores assigned by male and especially female reviewers; although differences in the mean values are small (and not statistically significant), it is important to note that these differences in results come mainly from the diverse way in which female evaluators on average assess applications with male or female PIs.

6. SUMMARY AND CONCLUSIONS

In this paper, we implemented a survey field experiment in the Galician University research funding and evaluation agencies at the time in which evaluation of the real applications to

the call was taking place. The experiment focused on identifying the effects of the gender of the application's PI alone, and in interaction with another factor (the gender of the reviewer). The main purpose of the experiment was not so much to confirm theories (with additional empirical evidence) as to confirm or not (the existence of) the regularities regarding gender bias. The experiment was built as a hypothetical funding application evaluation form in a factorial survey, was implemented in the same period as the real evaluation was being performed, using the same evaluation criteria and weighting in the evaluation form, and with the real reviewers selected by the funding agency.

Regarding our first research question, we found only small differences in the scores of the randomly assigned applications whereby male PI scores are higher than female PI ones, but these differences were not statistically significant. Thus, the results do not allow the null hypothesis that there are no differences in scores to be rejected. Estimating the effect size, we found small orders of magnitude.

As regards the second research question, the differences in scoring arise mainly from how female evaluators assign the scores to female and male PIs, with better scores for the latter. Although the differences are not statistically significant, estimating the effect size of the differences for reviewers of different genders we have found that, for female reviewers, the difference in the effect size between female PI and male PI scores is around 0.5 standard deviations, and for male reviewers, the effect size of the difference is negligible (0.02 SD).

Observational evidence about the existence of gender bias in research funding has yielded contradictory results, partly due to conceptual imprecision, but also because of methodological and measurement shortcomings, in addition to contextual or sampling effects that preclude controlling for unobserved heterogeneity. As a consequence, this type of observational research has looked for the causes of effects, adopting an approach in which gender bias (unobserved) is often considered as the residual of the gender disparity not explained by the observed variables, including merit.

In contrast, the experimental literature to which our study aims to contribute introduces randomization and/or manipulation to search for the effects of causes and to confirm (or reject) the existence of gender differences in the assessment that could support the idea of bias against one gender or the other. The study by Moss-Racusin et al. (2012), albeit analyzing a hiring decision, shares some methodological similarities with ours. In contrast with their results, we did not find that male applicants were rated significantly higher than female ones. In this sense, our findings are more in line with previous results (Carlsson et al., 2021; Ceci & Williams, 2015; Forscher et al., 2019; Williams & Ceci, 2015), which claim that there is not systematic bias against women in peer review evaluation. Nor did we find in our case that women were favored.

More recent studies based on German factorial surveys find that the evaluation favored women (Solga et al., 2023), and attribute the finding to the potential effect of "gender equality policies" in hiring procedures.

However, academic hiring is different from research funding, and with our data from the assessment of a set of highly experienced reviewers selected by the Galician Regional evaluation agency, only small differences were found; this connects with Ceci's claim (Ceci, 2018) about the relevance of the experience factor in research evaluation.

Previous experimental research has often been criticized for the lack of genuineness of the laboratory context, the nonauthenticity of the treatments, or the unrealistic character of the subjects. Experimental studies in funding agencies are rare. We have added to the literature a field study with a research design that used blocked randomization of the treatments and got

a rate of response of 100% of the population of evaluators involved in an ongoing research funding process, thus providing more robust evidence in terms of causal inference.

Of course, the context of the regional agency, the practices of selection of evaluators, and the behavioral effects of more than 15 years of gender equality policies could affect the evaluation (and it is expected to affect it over time), but determining the effect of the regulation over time was not the research question guiding our experiment.

Our experiment's main empirical result is that the gender of the applicant could not be seen as the direct cause (as we believe that the randomization of the treatment if properly implemented is a direct response to the effect of the causes approach) of a higher or lower rating of the funding application. Therefore, we found no support for the effects of gender of the PI on peer review outcomes, at least with our group of real reviewers.

In much of the previous experimental research, the gender of the reviewing participants has either not been a focus in the analysis or, when it has been, no effects on the responses were reported (Forscher et al., 2019; Moss-Racusin et al., 2012; Solga et al., 2023).

In our case, female and male reviewers differ in their assessments; however, these effects are not in line with the matching hypothesis (or with the claim that reviewers hold same-gender preferences, or make gender-role congruity associations) (Jayasinghe et al., 2003; Marsh et al., 2009, 2011). Nevertheless, some of our results show that female reviewers give male PIs higher scores than female PIs, up to 0.5 SD; this leadership "bonus" in the scoring of male PIs could suggest that stereotypes about gender attributes associated with leadership roles may operate (Eagly & Karau, 2002), at least in the case of the female reviewers in our study. Empirically, the ratings of female reviewers of female PIs show a higher dispersion than the ratings of male PIs.

The results emerging from our experimental design are robust, but we acknowledge that the grounds of the differences in female reviewers' assessment of male and female PIs could result from other nonobservable attributes of our population, such as related evaluation experience or the disciplines or fields of research; the idea that more senior or excellent female academics may apply more demanding expectations in peer review has already been suggested by empirical findings elsewhere (Borsuk et al., 2009; Cruz-Castro & Sanz-Menéndez, 2021). Factors such as the distribution of reviewers among different disciplines and research fields could also influence the aggregate values of the assessment, based on the different ways in which they interpret or provide meaning to the description of the application included in the experiment.

The findings have some policy implications regarding the rationale for increasing the number of female reviewers on the panels as a way to increase female funding success rates or their ratings, in line with the empirical findings of previous research showing a lack of effect of the gender composition of committees on the number of successful female candidates (Bagues et al., 2017; Zinovyeva & Bagues, 2015) as policy action.

As for the caveats in our research, we should acknowledge the following. First, the evidence presented, albeit robust, is limited in terms of generalizability to a broader community of reviewers, mostly because of the small number of subjects in the experiment and the local context and assessment practices.

In fact, in the real contexts of small countries or regional funding agencies, the number of applications and reviewers is limited to provide statistical generalizability to the results; in our case, as all reviewers involved in the evaluation participated and there were no dropouts, the potential limitations for induction and robust internal validity were not present. Moreover, in favor of the case, we should note that the funding agency under study implements a policy of

reviewer identification and selection from outside the region, targeting experienced scholars and scientists across several Spanish research institutions. Nevertheless, larger numbers would be needed to qualify the results by scientific field; developing a large-scale strategy of replication of the experiment in other funding agencies could also be further research.

Second, the type of instrument (group funding) and the limited level of competitiveness of the call (with success rates higher than 60%) may have contributed to the outcomes, as there is some observational evidence that women tend to be disfavored in more competitive contexts (Ors, Palomino, & Peyrache, 2013).

Finally, we have analyzed a policy instrument that mentions in the call the commitment to gender equality in science and academia, and this may have had a moderating impact on the gender effect, possibly linked to socially desirable behavior or rational adaptation to a changing policy environment.

To sum up, more comparative research is needed in other funding agencies to control for context-specific factors of this and other types; in this regard, replication of the experiment in other funding agencies would strengthen the findings.

ACKNOWLEDGMENTS

We would like to acknowledge the written comments and feedback on an earlier draft received from Peter van den Besselaar, Ulf Sandström, Fernando Galindo-Rueda, and Luis Miller, and also to the three anonymous QSS referees for their suggestions to improve the manuscript. The paper was also presented at the 8th Annual Conference of the Society for Economic Measurement (SEM 2023), and we thank the participants in our session and especially Catalina Martínez, the session organizer, for their comments. Special thanks to the officers (Faustino Infante in the funding agency (SXU) and Ana López in the evaluation unit (AGSUG)) who have provided support for the project. We also acknowledge the support of Sara Varela, from IMOP Insights, who was in charge of the implementation of the survey.

AUTHOR CONTRIBUTIONS

Laura Cruz-Castro: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Writing—original draft, Writing—review & editing. Luis Sanz-Menéndez: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

DATA AVAILABILITY

Original data supporting this study cannot be made publicly available due to clauses established in the research contract signed by SXU and CSIC. Anonymized data used in this paper could be provided, after consultation with the funders, if reasonable and justified to the authors. After finishing the publication of results and with the agreement of the funders, anonymized data supporting this study will be available at DIGITAL.CSIC (https://digital.csic.es/).

ETHICS STATEMENT

The authors' relationship with the funding (SXU) and evaluation (ACSUG) agencies was regulated by a research contract between the funding agency and the authors' employing institution (CSIC). The contract established the ethical procedures, including the transfer of personal data (name and email) for the purpose of the survey. The contract involves various relevant clauses from the ethical side. First, there was a general nondisclosure confidentiality agreement (regarding the diffusion of the agency name in any publication without their consent, a consent that was received in February 2023). The second one relates to the treatment of personal data of the evaluators involved (in fact only names and gender) that were provided for the project, and for which the contract stated that, if anonymized, they could be used in aggregate ways for research publications. The third referred to the commitment of the project team regarding the implementation of all ethical standards defined by the CSIC Ethics Committee.

The main ethical protocols and practices of relations with the evaluators were those of the evaluation agency (ACSUG) which was also responsible for the ethical clearance of the evaluation procedure and the survey. The CSIC ethical committee did not have any specific role, other than general guidance and principal definitions on personal data management, as was already established in the research contract. For more information about the ACSUG's code of ethics see: https://www.acsug.es/gl/documentacion/publicacions/c%C3%B3digo-%C3%A9tico (accessed on April 16, 2023).

FUNDING INFORMATION

The concept design for the experiment in a funding agency was developed in the context of the H2020 GRANteD project (grant agreement No. 824574). The research was funded by a research contract between the Regional Government of Galicia General Secretariat for Universities (SXU) and the CSIC Institute of Public Goods and Policies (reference: XG-CCEU-SXU-03/2022 and BDC: 20225046). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders or our employing institution.

REFERENCES

Aksnes, D. W., Piro, F. N., & Rørstad, K. (2019). Gender gaps in international research collaboration: A bibliometric approach. *Scientometrics*, *120*(2), 747–774. https://doi.org/10.1007/s11192-019-03155-3

Albers, C. J. (2015). Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences*, *112*(50), E6828–E6829. https://doi.org/10.1073/pnas.1518936112, PubMed: 26635232

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement*, *66*(6), 945–960. https://doi.org/10.1177/0013164406288161

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE Publications. https://uk.sagepub.com/en-gb/eur/factorial-survey-experiments/book240309. https://doi.org/10.4135/9781483398075

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, *107*(4), 1207–1238. https://doi.org/10.1257/aer.20151211

Barrera, D., Gerxhäni, K., Kittel, B., & Miller, L. (2023). *Experimental sociology: An outline of a scientific field* (forthcoming). Cambridge University Press.

Bello, M., & Galindo-Rueda, F. (2020). The 2018 OECD international survey of scientific authors. *OECD Science, Technology and Industry Working Papers 4/2020*. https://doi.org/10.1787/18d3bf19-en

Bendiscioli, S., Firpo, T., Bravo-Biosca, A., Czibor, E., Garfinkel, M., … Woods, H. B. (2022). *The experimental research funder's handbook*. Research on Research Institute. https://doi.org/10.6084/m9.figshare.19459328.v2

Bianchini, S., Llerena, P., Öcalan-Özel, S., & Özel, E. (2022). Gender diversity of research consortia contributes to funding decisions in a multi-stage grant peer-review process. *Humanities and Social Sciences Communications*, *9*, 195. https://doi.org/10.1057/s41599-022-01204-6

Bloch, C., Graversen, E. K., & Pedersen, H. S. (2014). Competitive research grants and their impact on career performance. *Minerva*, *52*(1), 77–96. https://doi.org/10.1007/s11024-014-9247-0

Bol, T., de Vaan, M., & van de Rijt, A. (2022). Gender-equal funding rates conceal unequal evaluations. *Research Policy*, *51*(1), 104399. https://doi.org/10.1016/j.respol.2021.104399

Bornmann, L. (2015). Interrater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for*

Information Science and Technology, *66*(12), 2415–2426. https://doi.org/10.1002/asi.23334
Bornmann, L., Mutz, R., & Daniel, H. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238. https://doi.org/10.1016/j.joi.2007.03.001
Borsuk, R. M., Aarssen, L. W., Budden, A. E., Koricheva, J., Leimu, R., ... Lortie, C. J. (2009). To name or not to name: The effect of changing author gender on peer review. *BioScience*, *59*(11), 985–989. https://doi.org/10.1525/bio.2009.59.11.10
Carlsson, M., Finseraas, H., Midtbøen, A. H., & Rafnsdóttir, G. L. (2021). Gender bias in academic recruitment? Evidence from a survey experiment in the Nordic region. *European Sociological Review*, *37*(3), 399–410. https://doi.org/10.1093/esr/jcaa050
Ceci, S. J. (2018). Women in academic science: Experimental findings from hiring studies. *Educational Psychologist*, *53*(1), 22–41. https://doi.org/10.1080/00461520.2017.1396462
Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236
Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, *108*(8), 3157–3162. https://doi.org/10.1073/pnas.1014871108, PubMed: 21300892
Ceci, S. J., & Williams, W. M. (2015). Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men. *Frontiers in Psychology*, *6*, 1532. https://doi.org/10.3389/fpsyg.2015.01532, PubMed: 26539132
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). London: Routledge. https://doi.org/10.4324/9780203771587
Cole, J. R. (1979). *Fair science: Women in the scientific community*. New York: The Free Press.
Cole, J. R., & Cole, S. (1979). Which researcher will get the grant? *Nature*, *279*(5714), 575–576. https://doi.org/10.1038/279575a0, PubMed: 450106
Cole, J. R., & Cole, S. (1981). *Peer review in the National Science Foundation: Phase II*. Washington, DC: National Academy Press. https://www.columbia.edu/cu/univprof/jcole/Phase2.pdf
Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, *214*(4523), 881–886. https://doi.org/10.1126/science.7302566, PubMed: 7302566
Cole, S., Rubin, L., & Cole, J. R. (1977). Peer review and the support of science. *Scientific American*, *237*(4), 34–41. https://doi.org/10.1038/scientificamerican1077-34, PubMed: 905818
Cole, S., Rubin, L., & Cole, J. R. (1978). *Peer review in the National Science Foundation: Phase one of a study*. Washington, DC: The National Academies Press. https://doi.org/10.17226/20041
Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally & Co.
Cox, D. R. (2020). Statistical significance. *Annual Review of Statistics and its Application*, *7*(1), 1–10. https://doi.org/10.1146/annurev-statistics-031219-041051
Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474. https://doi.org/10.1257/jel.47.2.448
Cruz-Castro, L., Ginther, D. K., & Sanz-Menéndez, L. (2023). Gender and underrepresented minorities differences in research funding. In B. Lepori, D. Hicks, & B. Jongbloed (Eds.), *Handbook of public funding of research* (pp. 279–300). Chichester: Edward Elgar. https://www.elgaronline.com/display/book/9781800883086/book-part-9781800883086-25.xml. https://doi.org/10.4337/9781800883086.00025
Cruz-Castro, L., & Sanz-Menéndez, L. (2020). *Grant allocation disparities from a gender perspective: Literature review. Synthesis report*. https://doi.org/10.20350/digitalCSIC/10548
Cruz-Castro, L., & Sanz-Menéndez, 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
Dawid, A. P., & Musio, M. (2022). Effects of causes and causes of effects. *Annual Review of Statistics and Its Application*, *9*(1), 261–287. https://doi.org/10.1146/annurev-statistics-070121-061120
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005, PubMed: 29331519
Derrick, G. E. (2019). *The evaluators' eye—Impact assessment and academic peer review*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-63627-6
Directorate-General for Research and Innovation (European Commission). (2021). *She figures 2021: Gender in research and innovation: Statistics and indicators*. Publications Office of the European Union. https://doi.org/10.2777/06090
Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573, PubMed: 12088246
Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 91–122. https://doi.org/10.1146/annurev-orgpsych-041015-062400
Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*(1), 275–298. https://doi.org/10.1146/annurev-psych-122216-011719, PubMed: 28961059
Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878, PubMed: 12051578
Forscher, P. S., Cox, W. T. L., Brauer, M., & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature Human Behaviour*, *3*, 257–264. https://doi.org/10.1038/s41562-018-0517-y, PubMed: 30953009
Fox, C. W., & Paine, C. E. T. (2019). Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, *9*(6), 3599–3619. https://doi.org/10.1002/ece3.4993, PubMed: 30962913
Fraser, D. A. S. (2017). *p*-Values: The insight to modern statistical inference. *Annual Review of Statistics and Its Application*, *4*(1), 1–14. https://doi.org/10.1146/annurev-statistics-060116-054139
Gaughan, M., & Bozeman, B. (2016). Using the prisms of gender and rank to interpret research collaboration power dynamics. *Social Studies of Science*, *46*(4), 536–558. https://doi.org/10.1177/0306312716652249, PubMed: 28948875
Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: W. W. Norton & Company.
Gërxhani, K., & Miller, L. (2022). Experimental sociology. In K. Gërxhani, D. De Graad, & W. Raub (Eds.), *Handbook of sociological science: Contributions to rigorous sociology* (pp. 309–323). Chichester: Edward Elgar. https://doi.org/10.4337/9781789909432.00026
Ginther, D. K., Basner, J., Jensen, U., Schnell, J., Kington, R., & Schaffer, W. T. (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PLOS ONE*, *13*(11),

e0205929. https://doi.org/10.1371/journal.pone.0205929, PubMed: 30427864
Hansen, J. A., & Tummers, L. (2020). A systematic review of field experiments in public administration. *Public Administration Review*, *80*(6), 921–931. https://doi.org/10.1111/puar.13181
Hug, S. E. (2022). Towards theorizing peer review. *Quantitative Science Studies*, *3*(3), 815–831. https://doi.org/10.1162/qss_a_00195
Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581, PubMed: 16173891
Imbens, G., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751
Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *166*(3), 279–300. https://doi.org/10.1111/1467-985X.00278
Jerrim, J., & Vries, R. (2023). Are peer reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications. *Social Science Journal*, *60*(1), 91–109. https://doi.org/10.1080/03623319.2020.1728506
Kahn, S., & Ginther, D. K. (2018). Women and science, technology, engineering, and mathematics (STEM): Are differences in education and careers due to stereotypes, interests, or family? In S. L. Averett, L. M. Argys, & S. D. Hoffman (Eds.), *The Oxford handbook of women and the economy* (pp. 767–798). https://doi.org/10.1093/oxfordhb/9780190628963.013.13
Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69. https://doi.org/10.1177/0013164404264850
Knight, C. R., & Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 275–299). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_14
Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128–161. https://doi.org/10.1037/a0036734, PubMed: 24865576
Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9
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. https://doi.org/10.1111/joes.12395
Larregue, J., & Nielsen, M. W. (2023). Knowledge hierarchies and gender disparities in social science funding. *Sociology*, 00380385231163071. https://doi.org/10.1177/00380385231163071
Lawrence, B. S., & Shah, N. P. (2020). Homophily: Measures and meaning. *Academy of Management Annals*, *14*(2), 513–597. https://doi.org/10.5465/annals.2018.0147
Leahey, E. (2006). Gender differences in productivity: Research specialization as a missing link. *Gender & Society*, *20*(6), 754–780. https://doi.org/10.1177/0891243206293030
Leahey, E. (2007). Not by productivity alone: How visibility and specialization contribute to academic earnings. *American Sociological Review*, *72*(4), 533–561. https://doi.org/10.1177/000312240707200403
Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. *Philosophy of Science*, *79*(5), 859–870. https://doi.org/10.1086/667841
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
Leeper, T. J. (2017). How does treatment self-selection affect inferences about political communication? *Journal of Experimental Political Science*, *4*(1), 21–33. https://doi.org/10.1017/XPS.2017.1
Levy, J., & Kimura, D. (2009). Women, men and the sciences. In C. Hoff Sommers (Ed.), *The science on women and science* (pp. 202–266). American Enterprise Institute.
Ley, T. J., & Hamilton, B. H. (2008). The gender gap in NIH grant applications. *Science*, *322*(5907), 1472–1474. https://doi.org/10.1126/science.1165878, PubMed: 19056961
Lloyd, M. E. (1990). Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis*, *23*(4), 539–543. https://doi.org/10.1901/jaba.1990.23-539, PubMed: 16795738
Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*, *79*(3), 1290–1326. https://doi.org/10.3102/0034654309334143
Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2011). Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model. *Journal of Informetrics*, *5*(1), 167–180. https://doi.org/10.1016/j.joi.2010.10.004
Mayer, S. J., & Rathmann, J. M. K. (2018). How does research productivity relate to gender? Analyzing gender differences for multiple publication dimensions. *Scientometrics*, *117*(3), 1663–1693. https://doi.org/10.1007/s11192-018-2933-1
McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444. https://doi.org/10.1146/annurev.soc.27.1.415
Mom, C., & van den Besselaar, P. (2022). Do interests affect grant application success? The role of organizational proximity. *arXiv:2206.03255*. https://doi.org/10.48550/arXiv.2206.03255
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109, PubMed: 22988126
Murray, D., Siler, K., Larivière, V., Chan, W. M., Collings, A. M., ... Sugimoto, C. R. (2019). Author-reviewer homophily in peer review. *bioRxiv*. https://doi.org/10.1101/400515
Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/titles/9620.html. https://doi.org/10.23943/princeton/9780691144511.001.0001
Nakamura, R. K., Mann, L. S., Lindner, M. D., Braithwaite, J., Chen, M.-C., ... Reed, B. (2021). An experimental test of the effects of redacting grant applicant identifiers on peer review outcomes. *eLife*, *10*, e71368. https://doi.org/10.7554/eLife.71368, PubMed: 34665132
Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, *3*(1), 601–630. https://doi.org/10.1146/annurev-economics-111809-125122
Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics*, *31*(3), 443–499. https://doi.org/10.1086/669331

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles*, *9*(3), 387–390. https://doi.org/10.1007/BF00289673

Pinholster, G. (2016). Journals and funders confront implicit bias in peer review. *Science*, *352*(6289), 1067–1068. https://doi.org/10.1126/science.352.6289.1067

Pohlhaus, J. R., Jiang, H., Wagner, R. M., Schaffer, W. T., & Pinn, V. W. (2011). Sex differences in application, success, and funding rates for NIH extramural programs. *Academic Medicine*, *86*(6), 759–767. https://doi.org/10.1097/ACM.0b013e31821836ff, PubMed: 21512358

Rayzberg, M. S. (2019). Fairness in the field: The ethics of resource allocation in randomized controlled field experiments. *Science, Technology, & Human Values*, *44*(3), 371–398. https://doi.org/10.1177/0162243918798471

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674982697

Sandström, U., & Hällsten, M. (2008). Persistent nepotism in peer-review. *Scientometrics*, *74*(2), 175–189. https://doi.org/10.1007/s11192-008-0211-3

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, PubMed: 33041361

Severin, A., Martins, J., Heyard, R., Delavy, F., Jorstad, A., & Egger, M. (2020). Gender and other potential biases in peer review: Cross-sectional analysis of 38 250 external peer review reports. *BMJ Open*, *10*(8), e035058. https://doi.org/10.1136/bmjopen-2019-035058, PubMed: 32819934

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-*N* design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8, PubMed: 29557067

Solga, H., Rusconi, A., & Netz, N. (2023). Professors' gender biases in assessing applicants for professorships. *European Sociological Review*, jcad007. https://doi.org/10.1093/esr/jcad007

StataCorp. (2021). *STATA user's guide*, Release 17. StataCorp LLC. https://www.stata.com/bookstore/users-guide/

Stewart, A. J., & Valian, V. (2018). *An inclusive academy: Achieving diversity and excellence*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/books/inclusive-academy. https://doi.org/10.7551/mitpress/9766.001.0001

Suarez, D., Fiorentin, F., & Pereira, M. (2023). Observable and unobservable causes of the gender gap in S&T funding for young researchers. *Science and Public Policy*, scad008. https://doi.org/10.1093/scipol/scad008

Swim, J., Borgida, E., Maruyama, G., & Myers, D. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Review*, *105*(3), 409–429. https://doi.org/10.1037/0033-2909.105.3.409

Titiunik, R. (2021). Natural experiments. In D. P. Green & J. N. Druckman (Eds.), *Advances in experimental political science* (pp. 103–129). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108777919.008

Traag, V. A., & Waltman, L. (2022). Causal foundations of bias, disparity and fairness. *arXiv:2207.13665*. https://doi.org/10.48550/arXiv.2207.13665

Treviño, L. J., Gomez-Mejia, L. R., Balkin, D. B., & Mixon, F. G. (2018). Meritocracies or masculinities? The differential allocation of named professorships by gender in the academy. *Journal of Management*, *44*(3), 972–1000. https://doi.org/10.1177/0149206315599216

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, PubMed: 17835457

van den Besselaar, P., & Mom, C. (2022). Gender differences in research grant allocation—A mixed picture. *arXiv:2205.13641*. https://doi.org/10.48550/arXiv.2205.13641

van den Besselaar, P., Mom, C., Cruz-Castro, L., Sanz-Menéndez, L., Hornbostel, S., & Möller, T. (2020). *Identifying gender bias and it causes and effects*. https://www.granted-project.eu/wp-content/uploads/2021/04/GRANteD_D2.1.pdf

van der Lee, R., & Ellemers, N. (2015a). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences*, *112*(40), 12349–12353. https://doi.org/10.1073/pnas.1510159112, PubMed: 26392544

van der Lee, R., & Ellemers, N. (2015b). Reply to Albers: Acceptance of empirical evidence for gender disparities in Dutch research funding. *Proceedings of the National Academy of Sciences*, *112*(50), E6830. https://doi.org/10.1073/pnas.1521336112, PubMed: 26635231

van der Lee, R., & Ellemers, N. (2015c). Reply to Volker and Steenbeek: Multiple indicators point toward gender disparities in grant funding success in The Netherlands. *Proceedings of the National Academy of Sciences*, *112*(51), E7038. https://doi.org/10.1073/pnas.1521331112, PubMed: 26647178

Volker, B., & Steenbeek, W. (2015). No evidence that gender contributes to personal research funding success in The Netherlands: A reaction to van der Lee and Ellemers. *Proceedings of the National Academy of Sciences*, *112*(51), E7036–E7037. https://doi.org/10.1073/pnas.1519046112, PubMed: 26647179

Walzenbach, S. (2019). Hiding sensitive topics by design? An experiment on the reduction of social desirability bias in factorial surveys. *Survey Research Methods*, *13*(1), 103–121. https://doi.org/10.18148/srm/2019.v13i1.7243

Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x, PubMed: 28458499

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *387*, 341–343. https://doi.org/10.1038/387341a0, PubMed: 9163412

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, *112*(17), 5360–5365. https://doi.org/10.1073/pnas.1418878112, PubMed: 25870272

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet*, *393*(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4, PubMed: 30739688

Zinovyeva, N., & Bagues, M. (2015). The role of connections in academic promotions. *American Economic Journal: Applied Economics*, *7*(2), 264–292. https://doi.org/10.1257/app.20120337