Experts on Scientific Committees

Milan Makany Natalia Zinovyeva*

August 11, 2025

Abstract

We study how the presence of experts affects decision-making in committees. Leveraging a natural experiment in Italy's national academic promotion system – where evaluators are randomly assigned to promotion committees – we find that better-published researchers apply significantly stricter standards and place greater weight on research quality relative to quantity. They also select candidates who have higher subsequent research productivity and promotion success. We show that expert evaluators not only influence outcomes directly, but also through their peers: other committee members put in more effort and become more demanding when paired with a high-quality colleague. However, we also find that committee service imposes substantial costs on the experts themselves. Serving on a two-year evaluation committee reduces a top researcher's own output by about 20–25% of an average year's publications, lowering their future willingness to serve.

JEL Codes: I23, J45, M54, D78

Keywords: committees, academic promotions, expert evaluators, evaluation quality, decision-making

*Milan Makany, Erasmus University Rotterdam & Tinbergen Institute, email: makany@ese.eur.nl; Natalia Zinovyeva, University of Warwick, email: natalia.zinovyeva@warwick.ac.uk. We would like to thank for their valuable comments Manuel Bagues, Anne Boring, Josse Delfgaauw, Jacques Mairesse, Otto Swank, as well as participants of seminars and talks at the Workshop on the Organisation, Economics and Policy of Scientific Research; European Economic Association; University of Edinburgh; University of Warwick; Erasmus School of Economics; University of Ottawa; Aalto University; SOFI; and CEBI at the University of Copenhagen.

1 Introduction

Committees play a central role in shaping outcomes across a wide range of institutions, from company boards, to central banks, to scientific boards and promotion panels. These collective decision-making bodies aim to pool information and reduce individual biases in high-stakes decisions. Experts – individuals with superior relevant knowledge or track records – are widely presumed to improve committee decisions given their experience and judgment. A vast theoretical literature analyzes how experts on committees may influence deliberation processes and committee decisions, under different procedural rules, strategic effort optimization, and reputation concerns (see, e.g., Ottaviani and Sørensen (2001); Levy (2007); Swank and Visser (2023)). However, there is relatively little empirical evidence documenting the impact of experts on committee dynamics and decisions. Even less is known about the mechanisms through which experts exert influence – whether by applying different standards, shaping deliberation, or affecting the behavior of their peers. This paper addresses these questions by analyzing a unique setting with random assignment of evaluators to committees, allowing us to cleanly identify the impact of evaluator expertise on committee outcomes.

Our empirical evidence is based on rich administrative data from Italy's Abilitazione Scientifica Nazionale (ASN), a centralized academic promotion system introduced in 2012 to combat nepotism and promote merit-based hiring. In the ASN, candidates for Associate or Full Professor at public universities need to receive a qualification granted by discipline-level 5-member committee before applying for promotion at the university level. In the ASN, evaluators are randomly selected from a pool of eligible professors who volunteer, and each committee serves for a two-year term. This random assignment creates a natural experiment: some committees, by chance, end up with evaluators who are top researchers in their field (we refer to these as "expert" evaluators), while others are composed of more average members. We use this random variation in committee composition to examine three interrelated questions: (i) how evaluator expertise affects the stringency, the criteria, and the quality of promotion decisions as revealed by future performance of promoted candidates, (ii) how expert presence influences the behavior of other committee members, and (iii) what

costs (in terms of foregone research output and future service) are borne by the experts themselves. We also explore whether experiencing these costs deters the most qualified individuals from volunteering in the future.

Exploiting the random assignment of evaluators to committees, we find that evaluators with stronger publication records apply stricter standards and place greater weight on publication quality relative to quantity. Replacing an average evaluator with one whose pre-committee publication record is one standard deviation stronger reduces a candidate's probability of success by 1.3 percentage points (3.4% of the 37% baseline success rate). At the same time, the marginal effect of normalized publication quantity declines by 0.6 percentage points (5.5% of the 11 percentage point baseline effect), while the marginal effect of normalized publication quality rises by 0.8 percentage points (7.3% of the 10 percentage point baseline effect).

We also find that expert involvement improves the outcomes of committee decisions, as measured by the future performance of the candidates selected. Candidates who are approved by higher-quality committees go on to produce more and better research in subsequent years. Using decade-long follow-up data, we show that a one-standard-deviation increase in committee evaluators' quality leads to an increase of about 3-4% of a standard deviation in the selected candidates' post-promotion citations. These candidates are also more likely to attain further career milestones: for instance, among those qualifying for Associate Professor, those evaluated by committees with more experts have a significantly higher probability of becoming full professors within the next 10 years. These differences persist even after controlling for a host of candidate observable characteristics (prior productivity, affiliation, etc.). This suggests that expert evaluators are better at recognizing "latent" potential in candidates.

Beyond their direct impact on committee decisions, expert evaluators influence the dynamics within committees. Using the random pairings of evaluators within committees, we show that committee members assigned to sit on a committee with better-published colleagues devote more effort to evaluation, become less lenient, and weigh publication quality more heavily. In particular, when an average evaluator is placed on a committee where one of the other members' average publication record is one standard deviation above what was

expected, she increases the length of the evaluation report by 7% of a standard deviation, reduces the probability of casting a positive vote by 1.5 percentage points, and raises the likelihood of supporting candidates with one-standard-deviation higher publication quality by 0.9 percentage points. These results show that experts shape committee outcomes both directly through their own votes and indirectly by inducing less-published colleagues to devote more effort, adopt stricter standards, and place greater weight on publication quality.

Enforcing higher standards comes at a substantial personal cost to experts. Serving on an ASN committee entails evaluating hundreds of candidates over a two-year period, on top of regular duties. We find that being randomly assigned to a committee causes a significant drop in the researcher's research output during and after the service term. Using data on all potential evaluators, we estimate that committee service leads to a decline in research output of approximately 20-25% of researchers' typical annual publications, with the effect spread over several years. One might expect that top researchers, with greater resources or experience, could better absorb or compensate for the evaluation burden. However, point estimates suggest, if anything, a larger absolute decline for those with stronger publication records (though the difference is not statistically significant). After experiencing a term on a committee, highly published researchers also become substantially less likely to volunteer again for future committees. The aggregate trends also show that top researchers become less likely to volunteer for committee service over time. This selective attrition of experts means that over time the pool of evaluators could become skewed away from the most qualified. In other words, there is a risk of adverse selection in who participates: the very people whose involvement most improves evaluation quality are those most deterred by the personal costs.

This paper makes several contributions to the literature. First, we extend research on how evaluator characteristics shape committee decisions. While prior work has examined the roles of gender (De Paola and Scoppa, 2015; Bagues et al., 2017; Card et al., 2020), committee experience (Seeber et al., 2021; Alexander et al., 2023), and academic proximity (Brogaard et al., 2014; Zinovyeva and Bagues, 2015; Li, 2017; Bagues et al., 2019a; Krieger et al., 2023), much less is known about how expertise affects evaluation. Existing evidence suggests that evaluators with weaker backgrounds may rely on noisier signals of quality (Bagues et al., 2019b). We provide the first causal estimate that better-published evaluators

systematically raise the bar for candidates and change the weighting of quality vs. quantity.

Second, we connect the literature on committee decision-making to research on peer effects in professional settings. Our results show that interactions among evaluators shape effort, evaluation criteria, and voting behavior, thereby influencing collective outcomes. This evidence also speaks to theoretical models of committee behavior, such as Ottaviani and Sørensen (2001) and the model of internal reputations developed by Swank and Visser (2023). They predict that the presence of better-informed or higher-quality peers raises each member's return to effort through within-committee reputation concerns. We provide the first causal micro-evidence consistent with these predictions: evaluators exert more effort when serving alongside top researchers, with the effect especially pronounced for strong scholars themselves. Importantly, these shifts in behavior are accompanied by higher-quality committee decisions, underscoring the broader relevance of this mechanism for understanding how peer composition affects institutional outcomes.

Third, we advance the understanding of the opportunity costs of academic service. Although it is widely recognized that peer review and committee work require substantial time commitments (Vernos, 2013; Aczel et al., 2021), prior evidence relies primarily on self-reported time-use data. In theory, heavier evaluation burdens might be offset if researchers reallocated effort more efficiently, limiting productivity losses. By exploiting random assignment to evaluation duties, we provide the first causal evidence on their impact on research productivity. We show that the productivity shock is both large and persistent, and that even highly accomplished researchers are not insulated from these long-term costs, despite streamlining their evaluation efforts.

Fourth, our findings speak to debates about the sustainability of evaluation systems. Many countries have adopted national structures to standardize promotions and foster fairness, yet the long-term viability of these systems depends on the continued participation of highly qualified evaluators. We show that evaluation service can induce selective withdrawal by top researchers, creating the risk of adverse selection in committee composition. If such dynamics persist, evaluation standards may erode over time, not through formal rule changes but through gradual shifts in who serves.

Although our evidence comes from national academic promotion committees, the mech-

anisms we document — higher standards from experts, peer influence within committees, and selective withdrawal due to high opportunity costs — are likely to operate in other forms of peer review, including journal refereeing, editorial boards, and hiring committees. Our results thus speak more broadly to the design and long-run viability of expert-driven evaluation systems.

Finally, our findings contribute to debates in research evaluation and scientometrics about whether excessive reliance on bibliometric indicators compromises peer review quality (Bertocchi et al., 2015; Stephan et al., 2017; Heckman and Moktan, 2020). We show that experts select candidates who subsequently receive more citations and are more likely to achieve promotions, even conditional on pre-evaluation observables. This pattern suggests that experts are better at extracting signals about unobservable aspects of candidate profiles, leading to higher-quality selection decisions. Our results highlight a complementary risk to that emphasized in the bibliometrics debate: when evaluation tasks impose high opportunity costs and offer limited personal rewards, the most capable evaluators may be systematically discouraged from contributing. Such selective withdrawal can have important consequences for the quality and sustainability of expert-driven evaluation systems.

More broadly, our study relates to research on team decision-making in high-stakes expert settings (Chan, 2021) and to personnel economics on incentive structures and worker productivity (Lazear and Rosen, 1981; Hoffman and Stanton, 2024), including their application to academia (Azoulay et al., 2011; Checchi et al., 2021; Nieddu and Pandolfi, 2022; Myers and Tham, 2023). In both domains, we show that the composition of committees influences individual behavior, and also collective outcomes.

The rest of the paper is organized as follows. Section 2 describes the institutional setting of Italian academic promotions, and Section 3 outlines our data sources. Section 4 presents the empirical analysis, and Section 5 concludes.

2 Italian National Academic Qualifications

Since 2012, the process of promoting associate and full professors in Italian public universities has followed a two-stage procedure. Candidates seeking promotion must first obtain a posi-

tive evaluation from a National Qualification Committee (Abilitazione Scientifica Nazionale, ASN) in one of 184 scientific fields before applying for promotion at the university level.[1]

The first National Scientific Qualification evaluations were conducted between 2012 and 2014.[2] In summer 2012, the Ministry issued calls for evaluator nominations and candidate applications, allowing researchers to apply to multiple fields and ranks. The application deadline was November 2012. Once the pools of eligible evaluators and applicants were finalized, committee members were selected by random draw between late November 2012 and February 2013. Upon appointment, each committee published its general evaluation criteria online, after which candidates had two weeks to withdraw. Committees then received the final candidate lists and completed their evaluations, with most results published between March and December 2013. Evaluators served two consecutive rounds: the second application deadline was October 2013, with evaluations held throughout 2014.

Each committee responsible for these evaluations consists of five members.[3] Members are randomly selected from a pool of eligible professors who volunteer and meet specific research productivity criteria. In STEM+M fields (Science, Technology, Engineering, Mathematics, and Medicine) and psychology, eligibility is determined by minimum thresholds for publications in scientific journals, citations, and the H-index. In SSH fields (Social Sciences and Humanities), the criteria include a minimum number of journal articles, articles in high-quality journals, and books or book chapters. Although the draw is subject to multiple restrictions, the procedure is clear, and the draw is random conditional on two observable evaluator characteristics: affiliation and academic area.[4] If a selected evaluator resigns, an-

1. The Abilitazione Scientifica Nazionale (ASN) was introduced as part of the Gelmini Reform under Law 240/2010.

2. Official documents regulating the process are available at http://abilitazione.miur.it/public/index.php?lang=eng (retrieved July 2026).

3. In the 2012-2013 evaluation cycle, four members were full professors from Italian public universities, and one was a professor from a foreign university in an OECD country. Since the 2015 cycle, all committee members have been selected from Italian universities. This study focuses exclusively on researchers based in Italy.

4. The draw procedure for 2012 is explained by the Ministry in the following document: https://abilitazione.mur.gov.it/public/documenti/Modalita_sorteggio.pdf (Accessed on June 27, 2025). For later waves it is described in the following document: https://abilitazione.mur.gov.it/public/documenti/2016/Allegato_1_Procedura_sorteggio_commissioni_ASN_2016.pdf (Accessed on June 27, 2025). Randomization is conducted in a way that leaves little room for manipulation: eligible evaluators in each field are listed alphabetically and assigned sequential numbers; a single random sequence is then drawn and applied to select committee members across multiple fields.

other individual is randomly drawn from the list. Committee members serve for a two-year term and are barred from participating in national evaluations for three years after completing their term. Their work is conducted *pro bono*, with only minimal reductions in teaching duties (if any).[5]

The evaluation relies solely on the candidate's CV and up to ten submitted publications. Committee work involves several meetings to define evaluation criteria and assess applications. The number and format of these meetings, whether in-person or online, depend on the volume of applications. In the 2012 cycle, committees had full discretion in setting evaluation standards, though the Ministry recommended referencing three bibliometric indicators – the same as used for establishing the eligibility of evaluators – when assessing candidates. In subsequent cycles, minimum conditions based on these indicators were formally introduced, though committees retained discretion in defining additional criteria.

Decisions on each candidate are made through a qualified majority, requiring at least four out of five votes in favor. Evaluators are required to accompany each of their individual votes with a written evaluation. Additionally, the committee as a whole is required to draft a summary document outlining its final decision on each candidate. Following the evaluation process, the names of candidates and evaluators, individual evaluation reports, and the committee's final decisions are published online for transparency.

3 Data

3.1 Administrative data on Italian professors

We collect administrative records on all professors employed at Italian public universities between 2010 and 2023 from the Italian Ministry of Universities and Research.[6] These data allow us to construct a comprehensive panel of faculty employed during this period, institutional affiliation, department, area of research specialization, academic rank, and gender.

5. While, according to the regulation, evaluators can apply for a temporal teaching reduction, most of the time such reductions are not operationally possible. Teaching and administrative obligations cannot be easily transferred across faculty members. Consequently, evaluators have limited flexibility to shift effort away from these nonresearch duties when committee service arises.

6. The data are available on the Ministry's website: https://cercauniversita.mur.gov.it/php5/docenti/cerca.php, last accessed 15 November 2024.

The dataset covers approximately 90,000 professors across 95 public universities over the sample period.

3.2 National Qualification: assessment panels

We use administrative data on academics listed as eligible evaluators for the National Qualification Evaluations (ASN).[7] The dataset includes individuals initially drawn by lottery to serve on evaluation committees, as well as those who ultimately participated following any resignations or substitutions. It covers five ASN evaluation cycles conducted between 2012 and 2023.

Across the five waves of examinations, around 15,000 professors from Italian public universities volunteered and qualified to be included in the pool of eligible evaluators. Table A1 provides descriptive statistics of the draw procedure and the pool of potential evaluators. On average, each field's pool of eligible evaluators consists of 26 professors. Each committee consists of five members, who are initially appointed via random draw. In the 2012 wave, one out of the five evaluators was a researcher affiliated with a foreign institution. This rule was abolished in subsequent waves. Approximately 9% of the evaluators initially drawn resigned and were replaced by other randomly selected eligible evaluators. The median Full Professor in the pool has 81 publications, of which around 95% are articles. The remaining 5% represents books, chapters, and conference proceedings.

3.3 National Qualification: candidates

We use information on applicants pre-registered in the first round of the national scientific qualification in 2012.[8] More than 46,000 researchers submitted their applications, accounting for around 61% of assistant professors and 60% of associate professors in Italy.[9] One third of candidates registered in several fields (e.g: qualification to full professorship in Political Economy and qualification to full professorship in Applied Economics) or in different cat-

7. The Ministry maintains a dedicated website with detailed information about the ASN: https://abilitazione.mur.gov.it/public/index.php, last accessed in July 2025.

8. Candidate curricula vitae (CVs) are publicly available on the ASN website during the evaluation process. We draw on data for pre-registered 2012 applicants collected by Bagues et al. (2017).

9. This is based on our own calculations using information from the Italian Ministry of Education on the identity of all assistant (*Ricercatori*) and associate professors (*Associati*) in Italy on December 31 2012.

egories of the same field (e.g.: qualification to full and associate professorships in Political Economy). In total, there were 69,020 pre-registered applications, approximately 375 per field.

Table A2 reports the descriptive statistics of candidates. Most candidates are already based at an Italian university, typically holding a permanent position in the field for which they seek promotion.[10] About one third of candidates are either unaffiliated with an Italian university or occupy non-tenure positions. On average, applicants have 16 years of research experience, measured from their first publication.

3.4 Sources of publication data

To measure research productivity of candidates and evaluators before and after the qualification exams, we combine data from three sources.

First, we extract candidates' publication records directly from the curricula vitae they submitted with their ASN applications. The advantage of this source is that it contains exactly the information available to evaluators when making their decisions, allowing us to accurately reconstruct candidates' observable research output at the time of evaluation.

Second, for both candidates and evaluators, we compile publication and citation data from two bibliometric databases. The primary source is IRIS (Institutional Research Information System), managed by ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca). IRIS comprises information on publications by researchers at all Italian public universities and research centers. Researchers individually report their publications through the database's online portal. The Ministry subsequently uses these records for national research assessments, which influence the allocation of research grants and contribute to university funding. The includes approximately eight million publications from 2000 to 2024.[11] We match 80% of all professors and 88% of evaluators to records in IRIS. To mitigate any biases from self-reporting, we supplement IRIS with OpenAlex, an open-access bibliometric index covering over 200 million works authored by approximately 90 million researchers worldwide (Priem et al., 2022). We successfully match 89% of professors in the

10. 42% of associate professor applicants hold tenured assistant professorships.
11. We collected information from universities' IRIS portals in December 2024.

Italian administrative database and 90% of evaluators to OpenAlex records.

We merge IRIS and OpenAlex data by de-duplicating on title, publication year, and journal. Whenever a publication appears in both sources, we retain its citation count from OpenAlex, since those counts are updated more frequently and cover more recent work. We link administrative records to publication data by matching on researchers' names, affiliations, and research fields. The full merge procedure is detailed in Appendix C.1. Together, IRIS and OpenAlex cover 96% of all Italian professors in our administrative dataset.

3.4.1 Candidate research profiles

We characterize candidates' research productivity along two dimensions, quantity and quality of publications. The quantity of publications reflects the volume and consistency of their research activity over time. The quality of publications proxies the impact and standing of their work within the scholarly community.

We measure quantity as the total number of academic publications listed on a candidate's CV. This measure includes journal articles, books, book chapters, conference proceedings, and patents. This broader set of outputs allows us to account for disciplinary differences in publication practices, particularly in SSH fields where books and chapters often constitute the primary forms of research dissemination.

We measure quality as the share of journal articles classified as high-impact. Constructing quality as a share allows us to avoid conflating it with the non-linear effects of publication quantity. We define high-impact publications differently across fields to account for disciplinary norms. In STEM+M fields, we classify as high-impact publications appearing in journals ranked in the top quartile (Q1) of the Web of Science based on the Article Influence Score.[12] In economics and business, as well as in other SSH fields, we use the list of high-impact journals (the so-called *A-journals*) compiled by the Italian evaluation agency, which includes approximately 7,000 academic journals.

While any journal-based metric imperfectly captures the intrinsic quality of individual publications, we rely on journal rankings because they offer an ex-ante signal of research

12. The Article Influence Score is an indicator similar to the 5-year Impact Factor, but it offers several advantages: it weights citations by the quality of the citing journal and by the inverse of the number of references in the citing journal, and it excludes self-citations.

quality that is observable to evaluators at the time of assessment. Relying on citation counts may introduce a risk of popularity biases or a risk of systematically understating the impact of more novel or innovative work, which often follows a delayed citation trajectory and accrues recognition over longer time horizons (Wang et al., 2017; Koffi, 2025). In contrast, journal rankings provide a better contemporaneous proxy for perceived scientific value. Furthermore, the use of journal-level indicators is well established in Italian academia as well as in other contexts of research evaluation.

We use the thresholds of Q1 journals in STEM+M fields and A-journals in SSH fields to ensure that our measure of publication quality is comparable across disciplines. While continuous metrics such as the Article Influence Score or the Impact Factor are widely available and informative in STEM+M fields, they are often unavailable or less consistently applied in many SSH disciplines. Relying on categorical classifications allows us to maintain consistency across fields and to align our measures with those used in national evaluation exercises and widely accepted disciplinary standards.

One limitation of our approach is that we do not systematically classify the quality of books, book chapters, or other non-journal outputs, which are particularly important in several SSH disciplines. As a result, our measure of publication quality primarily captures journal-based indicators and may understate or fail to capture the impact of high-quality books or edited volumes.

In STEM+M fields, the average share of publications per candidate appearing in top-quartile journals varies between 30% and 50%. In economics and business, the average share is approximately 15% for top-quartile journals and 21% for A-journals. In the remaining social sciences and humanities fields, the average share is about 29% for A-journals.

To account for structural differences in publication practices across disciplines, we normalize both the quantity and quality measures within each field. Specifically, we express each candidate's quantity and quality as standardized z-scores relative to candidates in the same discipline. This approach ensures that the variation we exploit reflects differences within fields rather than systematic differences across them.

We acknowledge that reducing all aspects of a candidate's multidimensional research output to a single measure of quantity and a single measure of quality is restrictive. Research

profiles encompass many dimensions that cannot be fully captured by any limited set of indicators, if they can be captured by objective criteria at all.[13] However, our aim is more modest: to assess whether committees with different characteristics place relatively more weight on these two widely observed, and widely cited, dimensions of academic performance when making promotion decisions. Focusing on quantity and quality makes it easy to see the key trade-off clearly and testable in the data. In Appendix B.2, we assess the robustness of our findings to alternative metrics of candidate quantity and quality.

3.4.2 Evaluator and committee profiles

For the purposes of our analysis, we construct a single measure of evaluator research quality designed to capture both research productivity and impact. Since both the quantity and quality of publications are important dimensions of research activity, we measure evaluator productivity as the total number of high-impact articles produced in the ten years preceding the exam. We normalize evaluators' research quality at the level of the relevant pool of potential evaluators.

We measure committee quality as the average of individual evaluators' standardized publication scores. This measure captures the committee's collective research strength and exploits the full variation in evaluator productivity, offering several practical advantages. First, the mean aggregates information from all members, reducing idiosyncratic noise and measurement error. Second, it aligns with the idea that each member contributes incrementally to setting evaluation standards, so the committee's average productivity maps to the intensity of scrutiny. Third, using the mean yields a continuous measure with a straightforward interpretation.

However, this choice may conceal heterogeneity in individual influence: a committee of uniformly average researchers and one combining very high- and very low–productivity evaluators can share the same mean, even though their deliberation dynamics might differ. For instance, highly productive members may help set the tone for evaluation and shape committee criteria. In Appendix B.3, we assess the robustness of our findings to alternative committee-quality metrics and show that our main results hold under these specifications.

13. For a detailed discussion of other dimensions of candidate research profiles, see Appendix B.1.

4 Empirical analysis

In this section, we examine the role of better-published researchers in shaping evaluation outcomes and the behavior of committee members. First, we analyze how committee composition influences evaluation criteria and selection standards. Second, we assess how the quality of the committee relates to the subsequent academic performance of selected candidates. Third, we estimate the effect of committee participation on the research productivity of evaluators. Fourth, we document how evaluators adjust their effort when preparing reports. Finally, we study how the presence of better researchers on a committee affects the behavior of their peers.

4.1 Committee quality and evaluation criteria

In this subsection, we investigate whether evaluation committees that include relatively better-published researchers apply stricter standards and base favorable judgments on a different set of candidate characteristics than other committees. In particular, we examine whether such committees weight publication quality and quantity differently. Given their experience with rigorous publication standards and their reputational stake in preserving those standards, we expect these committees both to apply more stringent criteria and to place greater emphasis on the quality of candidates' work.

We estimate the following equation on the sample of all candidates who submitted an application in 2012:

$$\begin{aligned} \textit{Qualified}_{ie} = \beta_0 + \beta_1 N_i + \beta_2 Q_i + \beta_3 T_e + \beta_4 T_e \times N_i + \beta_5 T_e \times Q_i \\ + \beta_6 \mathbb{E}\left[T_e\right] + \beta_7 \mathbb{E}\left[T_e\right] \times N_i + \beta_8 \mathbb{E}\left[T_e\right] \times Q_i + \varepsilon_{ie} \end{aligned} \tag{1}$$

where N_i denotes the total number of publications a candidate has produced in the previous ten years, and Q_i represents the proportion of high-impact publications over the same period. Committee quality, T_e, is defined as the average publication quality of all committee members, T_j, in exam e.[14] The expected committee quality, $\mathbb{E}[T_e]$, is computed by simulating one million times the official draw procedure, computing each evaluator's probability of

14. We discuss measurement in Sections 3.4.1 and 3.4.2. We perform robustness checks in Appendix B.

being selected $\mathbb{P}[drawn_{je}]$, and then weighting their T_j by these probabilities:

$$\mathbb{E}[T_e] = \sum_{j \in e} \mathbb{P}[drawn_{je}] \times T_j.$$

All candidate-level measures (N_i and Q_i) are standardized within each exam, and evaluator-level measures (T_j) are standardized within each pool of eligible evaluators. In equation 1, coefficients β_1 and β_2 capture the effect of a one-standard-deviation increase in publication quantity or quality on the probability of qualification; β_3 measures the causal impact of assigning a committee where all evaluators are one standard deviation better-published as compared to other potential evaluators in the field; and β_4 and β_5 show how higher committee quality changes the returns to publication quantity and quality.

Our identification strategy exploits the random draw of evaluators: any deviation of the realized committee composition from its expected composition, given the full pool of eligible evaluators, stems from lottery variation. For the randomization check see Appendix D. Given that some initially drawn evaluators resigned and were substituted by other randomly drawn evaluators, we instrument the final committee quality (and its interactions) with the initial-draw composition before any resignations took place. We estimate equation 1 by two-stage least squares (2SLS).

Table 1 reports the regression results. As expected, both the quantity and quality of publications are positively associated with a candidate's likelihood of qualification (column 1). A one–standard-deviation increase in the number of publications and in the share of high-impact articles is associated with a 11- and 10-percentage-point increase in the probability of qualification, respectively.

More strikingly, we find that higher-quality committees apply stricter standards: a one-standard-deviation increase in the quality of all evaluators is associated with an 5-percentage point reduction in the probability of qualification. Because committees consist of four Italian members, substituting one evaluator with another who is one standard deviation better published reduces an average candidate's success rate by 1.3 percentage points, or 3.4 percent relative to the 37 percent baseline.

The interactions indicate that higher-quality committees place greater weight on publica-

tion quality and less on quantity (column 2). When a candidate is evaluated by a committee whose members are, on average, one standard deviation better published, the marginal effect of publication quantity decreases by 2.5 percentage points, while the marginal effect of publication quality increases by 3 percentage points. In column 3, we control for candidate–evaluator connections to ensure the results are not driven by homophily. The findings are robust to controlling for coauthorship, shared affiliation, and research proximity.

Columns 4 and 5 split the sample between applicants to associate and full professorships, respectively. The increase in strictness and the reduced marginal returns to quantity are driven primarily by associate-level candidates, while the increased marginal returns to quality apply to candidates at both levels.

Columns 6 and 7 report results separately for STEM+M and SSH fields. Publication quantity is valued significantly more in STEM+M, whereas the marginal effect of publication quality does not differ across the two broad disciplinary groups. Although higher committee quality appears to tighten standards more in STEM+M and to amplify the emphasis on quality more in SSH, neither difference is statistically significant at conventional levels.

Next, we analyze whether increased strictness and the change in criteria is due to the direct voting behavior of better-published researchers or through their influence on criteria at the committee level. We estimate the following equation on the sample of all individual votes during the 2012 evaluations:

$$\begin{aligned} \textit{Positive vote}_{ije} = \beta_0 + \beta_1 T_j + \beta_2 T_{e,-j} + \beta_3 T_j \times T_{e,-j} \\ + \beta_4 \mathbb{E}\left[T_{e,-j}\right] + \beta_5 \mathbb{E}\left[T_{e,-j}\right] \times T_j + \boldsymbol{\gamma} X_i + \varepsilon_{ije} \end{aligned} \tag{2}$$

where T_j is the publication quality evaluator j, and $T_{e,-j}$ is the publication quality of all other evaluators on the same committee in exam e. Formally, for each evaluator assigned to a committee, we define the realized peer quality as:

$$T_{e,-j} = \frac{1}{N_e - 1} \sum_{\{k \neq j\} \in e} \textit{drawn}_{ke} \times Q_k$$

where Q_k is evaluator k's quality-adjusted publication record measured 10 years prior to the exam, $\textit{drawn}_{ke}$ indicates whether evaluator k was selected for committee e, and N_e is the

number of evaluators on committee e.

We compute the expected peer quality as:

$$\mathbb{E}\left[T_{e,-j}\right] = \frac{1}{N_e - 1} \sum_{\{k \neq j\} \in e} \mathbb{P}\left[drawn_{ke} \mid drawn_{je} = 1\right] \times Q_k$$

where $\mathbb{P}\left[drawn_{ke} \mid drawn_{je} = 1\right]$ is the conditional probability that evaluator k is selected for committee e, given that evaluator j is also selected. These probabilities are simulated one million times following the official draw procedure.

The coefficient β_1 captures whether better-published evaluators are systematically more or less likely to cast positive votes. A negative coefficient would suggest that such evaluators directly influence committee outcomes by applying stricter individual voting standards. The coefficient β_2 identifies the causal effect of being randomly paired with higher-quality peers on an evaluator's propensity to vote favorably. A negative estimate of β_2 would imply that better-published researchers also influence their colleagues' voting behavior, consistent with peer effects operating at the committee level. Finally, the interaction term β_3 tests whether the influence of peer quality depends on an evaluator's own publication record, capturing potential heterogeneity in the influence peer effects among high-quality researchers.

The results of this regression are reported in column 1 of Table 2. We find that better-published evaluators are, on average, stricter in their voting: a one–standard-deviation increase in an evaluator's publication record is associated with a 2.5 percentage point decrease in the likelihood of casting a positive vote. We also find strong evidence of peer effects. When an evaluator is randomly assigned to serve alongside a colleague who is one standard deviation better published, their probability of casting a positive vote falls by 1.5 percentage points.[15] Finally, we do not find evidence of heterogeneity in peer effects: all evaluators, regardless of their own publication record, become stricter when surrounded by higher-quality peers.

Columns 2 and 3 of Table 2 introduce interactions between evaluator, peer, and candidate characteristics to examine whether peers influence how evaluators weigh different aspects of a

15. Each committee consists of four Italian members, so each evaluator has three peers. The coefficient estimate on average peer quality is 0.046. Dividing by three yields the per-peer effect of approximately 1.5 percentage points.

candidate's profile. Consistent with earlier findings, we find that better-published evaluators place greater emphasis on publication quality: an evaluator who is one standard deviation better published is 0.8-0.9 percentage points more likely to cast a positive vote when evaluating a candidate with a one–standard-deviation larger share of high-impact publications.

At the same time, we find that peers also shape the weight evaluators assign to publication quality. When an evaluator is randomly assigned to serve with a peer who is one standard deviation better published, the marginal effect of a candidate's publication quality on the probability of receiving a positive vote increases by an additional 0.9 percentage points.

These results indicate that better-published evaluators affect evaluation criteria through two distinct channels. First, they apply stricter standards themselves and place greater weight on publication quality when assessing candidates. Second, they influence the behavior of their fellow committee members. Evaluators randomly assigned to serve alongside better-published peers also become more stringent in their voting and attribute higher weight to the quality of publications. These results suggest that expert evaluators play an active role in shaping committee criteria, not only by raising overall selectivity, but also by increasing the weight placed on research quality.

4.2 Committee quality and qualified candidates' outcomes

This tendency of expert evaluators to favor candidates with fewer, higher-quality publications may simply reflect different preferences for modes of scientific communication. Alternatively, it may signal that top researchers are better able to discern true research quality, resulting in higher success rates for applicants with high-impact work.

In this subsection, we test whether candidates promoted by randomly assigned, relatively better-published committees subsequently achieve higher research productivity and better career outcomes. We follow Zinovyeva and Bagues (2015), who demonstrate that the future performance of successful candidates can serve as a proxy for evaluation quality, capturing both observable and unobservable dimensions of candidate quality at the time of the exam. We estimate whether qualification by a committee with higher research quality predicts

greater future research productivity and career success:

$$y_{iet} = \beta_0 + \beta_1 T_e + \beta_2 \mathbb{E}\left[T_e\right] + X_i\gamma + \varepsilon_{iet} \tag{3}$$

where y_{iet} denotes the post-qualification outcome of candidate i evaluated by committee e in year t. The set of controls X_i include the candidates' pre-exam affiliation, academic rank, the number of publications, the share of high-impact articles, and citations. The coefficient β_1 captures whether candidates qualified by randomly better-published committees subsequently achieve higher productivity. Standard errors are clustered at the exam level.

In the previous section, we showed that better-published committees qualify fewer candidates. Raising the bar may yield a smaller pool of, on average, stronger candidates, but may not guarantee better selection. To isolate the selection-criteria channel, we divide candidates by whether their committee's random evaluator quality shock was above or below the median. Although 42 percent of candidates who were evaluated by better-published committees were promoted versus 49 percent for lower-quality committees, restricting the latter to unanimously promoted candidates (43 percent) equalizes admission rates. In this restricted sample, a larger β_1 in equation 3 would indicate that higher-quality committees select top candidates more effectively.

We estimate equation 3 by two-stage least squares (2SLS), instrumenting the final committee quality T_e with the quality of the initially assigned committee to address endogeneity from evaluator resignations.

Table 3 examines whether candidates qualified by higher-quality committees produce more high-impact research in the five years following the exam. Panel A presents estimates for the full sample of promoted candidates, and Panel B for the unanimity-restricted sample. In the full sample, a one-standard-deviation increase in evaluator quality implies that a top-ranked candidate will publish 6 percent of a standard deviation higher share of articles in high-impact journals (column 3) and receive 5.1 percent of a standard deviation more citations (column 5). In the unanimity-restricted sample, the point estimates are slightly smaller and less precise, but generally confirm that candidates promoted by better-published committees achieve greater post-exam productivity than equally ranked candidates from

other committees.

We also examine whether the higher post-exam productivity of candidates promoted by better-published committees can be fully explained by their observable characteristics at the time of evaluation. We control for candidates' pre-exam affiliation, academic rank, measures of publication quantity and quality, and citation received by the candidate's publications by the time of the evaluation. This allows us to assess whether better-published committees merely weight observable measures more effectively or whether they better identify talent beyond easily observable indicators.

We find that both in the full and restricted sample the number of publications (column 2) and share of high-impact publications (column 4) can be explained by observable pre-evaluation characteristics. However, higher post-exam citations of publications by candidates promoted by better-published committees cannot be fully accounted for by pre-exam research profiles. Being promoted by a committee with one-standard-deviation better-published evaluators implies 3.9 percent of a standard deviation more citations in the full sample and 3.3 percent of a standard deviation more citations in the restricted sample, conditional on observables.

In Figure 1, we assess whether candidates selected by higher-quality committees achieve promotions more quickly following the exam. The main panels use the unanimity-restricted sample; Appendix Figure A1 shows the same estimates for the full sample. In each panel, we report both unconditional estimates and estimates that control for pre-exam affiliation, rank, and research productivity.

Panel A considers top-ranked Associate Professor candidates. A one-standard-deviation increase in evaluator quality is associated with a 6 percentage-point higher probability of promotion within ten years (from a 67.8% baseline), which falls to 2.5 points once controls are added. Panel B repeats this exercise for top-ranked Full Professor candidates: the point estimates remain positive but are not statistically significant.

In Panel C, we return to the top-ranked Associate Professor sample and examine eventual promotion to Full Professor. There, a one-standard-deviation rise in evaluators' quality corresponds to a 5.4 percentage point increase in the ten-year promotion rate (from 22.3% at baseline) in the unconditional model. Once we condition on pre-exam observables, the

effect decreases to 3.9 percentage points implying that replacing an average evaluator with a peer who is one standard deviation better-published increases this probability by roughly 6%.

We interpret these findings as demonstrating that higher-quality committees systematically select candidates with superior long-run potential. A key assumption underlying this interpretation is that the decision of the committee affects candidates' future outcomes only through the selection of candidates with higher potential, rather than through any direct effect. Specifically, we assume that being qualified by a higher-quality committee does not convey additional information to journal editors, referees, or university-level promotion committees beyond the fact of qualification itself. Instead, if the identity or perceived quality of the evaluating committee directly influenced subsequent recognition or opportunities, our estimates would conflate the committee's selection with such signaling effects, making our estimates upward biased. However, national qualifications are awarded to a large share of applicants – rendering any single committee's endorsement relatively indistinctive. Panel C also provides strong support for the selection interpretation. It shows that committee quality at the Associate-Professor stage predicts eventual promotion to Full Professor more than a decade later, under an independent review process. The temporal and institutional separation between the qualification exam and the full professor evaluation effectively rules out a sustained signaling effect and confirms that better-published committees are particularly effective in identifying candidates with long-term scholarly promise.

The fact that these effects persist after conditioning on all pre-exam observables implies that high-quality committees do more than optimally weigh readily available productivity metrics. Rather, they appear better at discerning the latent dimensions of candidate potential that predict subsequent scientific impact and accelerated career advancement.

4.3 Cost of committee service

In this subsection, we estimate the productivity cost associated with serving on evaluation committees. Specifically, we compare the publication trajectories of evaluators randomly selected for committee service with those of similar researchers not selected.

We estimate the following regression using the sample of all potential evaluators partici-

pating in the Italian Qualification System between 2012 and 2023:

$$y_{jet} = \beta_0 + \beta_1 drawn_{je} + \beta_2 \mathbb{P}\left[drawn_{je}\right] + X_j\gamma + \mu_e + \varepsilon_{jet} \tag{4}$$

where y_{jet} denotes the research output of researcher j in year t after evaluation e. The variable $drawn_{je}$ is a dummy variable indicating whether the researcher was selected to serve on the evaluation committee. The coefficient β_1 captures the causal effect of committee service on subsequent research productivity. Our identification relies on random assignment, conditional on the probability of selection, $\mathbb{P}\left[drawn_{je}\right]$.

We control for researchers' pre-exam productivity trajectories, including the number of publications, the share of high-impact articles, and include exam fixed effects. All productivity measures are normalized at the exam-year level to account for differences across research fields. We instrument the actual committee participation with the initial random assignment to address potential bias from endogenous resignations.

Figure 2 shows that committee service reduces researchers' publication output by approximately 5% of a standard deviation over the subsequent three years. Given that the standard deviation of total research output is approximately 10 publications, this decline corresponds, on average, to 1.5 fewer published works over the three years. This effect is economically meaningful, suggesting nontrivial opportunity costs of service. This decline in publication output is likely due to the fact that time devoted to committee service crowds out time otherwise allocated to research activities. There is the possibility that researchers might simply work longer hours to absorb the additional workload. Although this margin is unobservable in our data, the decline in research output suggests that any such adjustments are insufficient to fully offset the costs of committee service.

Figure 3 presents evidence using a cumulative outcome measure y_{jet}, where publication counts are set to zero in the year of the exam and are accumulated afterwards. Committee service leads to a persistent decline in research output that remains statistically significant for up to four years following committee participation. Although researchers eventually recover, suggesting some projects may merely be delayed rather than abandoned, the coefficients remain consistently negative (albeit statistically not different from zero) even beyond four

years. This may indicate potential lasting disruption to researchers' productivity trajectories.

While the differences in estimated impacts between more- and less-productive researchers are not statistically significant, point estimates suggest that top researchers may bear greater productivity costs. As shown in Figure 4, the estimated effect for top researchers is approximately 7% of a standard deviation, compared to 4.3% among lower-productivity peers. This pattern highlights a key trade-off: researchers with stronger publication records may improve candidate selection, but may also face higher opportunity costs when serving on evaluation committees.

Interestingly, we find systematic differences in strategic adjustments. Better-published researchers appear to reallocate effort away from lower-impact projects in order to preserve output on higher-impact work. Figure 5 shows that those with higher baseline productivity incur a substantially larger and statistically different reduction in low-impact publications relative to their lower-productivity peers. In the first year following the evaluations, their lower-impact output declines by 11 percent of a standard deviation. Consistent with a reallocation of effort, we detect no statistically significant change in their high-impact publications. This strategic shift, however, entails a long-run cost. A decade after the evaluations, these researchers still produce 12 percent of a standard deviation fewer low-impact publications, corresponding to an average loss of 5.5 articles.

Figure 6 presents descriptive trends in participation rates over time. Top researchers were initially more likely to volunteer, conditional on eligibility. However, this pattern reversed following the first evaluation wave. Since then, their likelihood of volunteering has steadily declined, in both STEM+M and Social Sciences, despite continued eligibility.

Table 4 provides causal evidence that past service reduces the likelihood of future participation, especially among more productive researchers. On average, randomly drawn researchers are approximately 20 percentage points less likely to volunteer in subsequent waves. Moreover, those with pre-exam productivity one standard deviation above the mean are an additional 2 percentage points less likely to serve again.

We interpret these patterns as evidence of endogenous selection out of committee service among highly productive researchers. While better-published researchers are initially more likely to volunteer, experience with service may reveal its substantial costs. As a result, they

may become increasingly reluctant to participate in future rounds.

4.4 Adjustments in effort

In this subsection, we investigate whether evaluator quality is correlated with effort and committee-level dynamics in effort adjustment. First, we provide descriptive evidence on the relationship between evaluators' baseline research productivity and the amount of effort they devote to their evaluation reports. Because effort is not directly observable, we use textual features of the evaluation reports as proxies. These include the total number of words and an inverse-document-frequency (IDF) weighted word count, which capture both the length and lexical diversity of the text. Second, we examine whether the composition of committees matter for how much effort evaluators exert. Specifically, we empirically test predictions coming from the theoretical literature on internal reputation in committee decision making (Swank and Visser, 2023).

We provide descriptive evidence on differences in effort exerted by better-published researchers by estimating the following specification using all evaluation reports from the 2012 wave of the Italian Qualification System:

$$textual\ feature_{ije} = \beta_0 + \beta_1 T_j + \beta_2 N_i + \beta_3 Q_i + X_{ij}\gamma + \varepsilon_{ije} \tag{5}$$

The unit of observation is an evaluator-candidate pair. The outcome variable is a textual proxy for evaluator effort. The key independent variable, T_j, measures the pre-exam research productivity of evaluator j. We control for the candidate's research profile, including the number of publications N_i and the share of high-impact publications Q_i. The vector X_{ij} includes additional controls for evaluator-candidate connections, such as prior coauthorship, shared affiliation, or specialization in the same subfield.

We present the estimation results in Table 5. All textual measures are positively associated with candidate productivity and with proximity indicators between candidates and evaluators. This pattern suggests that evaluators tend to exert more effort when assessing collaborators, peers, or candidates within their own field. High-performing candidates may elicit more thorough assessments, as evaluators perceive them as more deserving of detailed

feedback. Familiarity with a candidate's work or area of expertise may also reduce the cognitive cost of evaluation, enabling more specific and substantive commentary.

In contrast, we find that for the same candidate more productive evaluators write shorter and less complex reports. A one standard deviation increase in evaluator research productivity is associated with a 5.3 percent of a standard deviation decrease in report length (column 2) and a 6.4 percentage points decrease in linguistic complexity. These findings are consistent with the idea that highly productive researchers face higher opportunity costs and therefore optimize by exerting less observable effort in the evaluation process. Nevertheless, they may still incur greater costs in absolute terms.

One interpretation is that top researchers possess an absolute advantage in evaluation: their assessments may yield higher informational content for any given level of effort. If the marginal productivity of effort is higher for experienced evaluators, shorter reports may still be of equal or greater quality. Another possibility is that the main contribution of top researchers occurs earlier in the evaluation process, during the committee's deliberation over evaluation criteria. In the first wave of Italian Qualifications, committees had discretion to define the standards for qualification. Senior or more accomplished researchers may play a disproportionate role in setting these standards, guiding discussions on what constitutes a meaningful contribution, and shaping the emphasis placed on specific candidate attributes. In this interpretation, the relatively lower observable effort of top researchers during individual candidate reviews may not imply disengagement, but rather reflect the front-loading of their contributions during the design of the evaluation process itself.

As we discuss later, there are additional issues with interpreting these descriptive results. Since we are not exploiting random variation in the descriptive setting, observed relationships between evaluator characteristics and effort may reflect confounding factors. Specifically, aggregation bias can arise when peer effects are incorrectly attributed to individual evaluators in pooled specifications.

We next examine how committee composition affects evaluator effort, in line with predictions from the theoretical literature. So far, our analysis has abstracted from committee-level dynamics, however, prior work suggests that the presence of capable peers can influence individual behavior. In particular, Swank and Visser (2023) predict that evaluators exert

greater effort when they anticipate their assessments will be scrutinized by more accomplished colleagues, in order to maintain professional standing and avoid reputational loss. This mechanism implies that evaluators should increase their effort when assigned to committees composed of higher-quality researchers.

We provide causal evidence for this hypothesis by exploiting the random assignment of evaluators to committees in the 2012 wave of the Italian Qualification System. Conditional on being selected, some evaluators are randomly paired with more distinguished peers, while others are assigned to less accomplished committees purely by chance. We estimate the following model on the sample of all evaluation reports in the 2012 wave of the Italian Qualification System:

$$\begin{aligned} \textit{textual feature}_{ije} = \beta_0 + \beta_1 T_j + \beta_2 T_{e,-j} + \beta_3 T_j \times T_{e,-j} + \beta_4 \mathbb{E}\left[T_{e,-j}\right] \\ + \beta_5 \mathbb{E}\left[T_{e,-j}\right] \times T_j + \beta_6 N_i + \beta_7 Q_i + X_{ij}\gamma + \varepsilon_{ije} \end{aligned} \tag{6}$$

where the dependent variables are textual features of evaluation reports. Peer quality, $T_{e,-j}$, is defined as the average publication quality of all other evaluators on the committee, excluding j. The coefficient β_2 captures the causal peer effect – whether evaluators exert more effort when assigned to committees with more productive colleagues. Conditional on the expected composition $\mathbb{E}\left[T_{e,-j}\right]$, variation in realized peer quality is driven purely by the random assignment. We also include candidate-level controls for publication quantity N_i and quality Q_i, as well as evaluator-candidate connection indicators X_{ij}.

The coefficient β_3 captures whether more productive evaluators respond differently to the quality of their peers. A positive β_3 would suggest that high-performing evaluators increase their effort more when surrounded by equally strong colleagues.

Table 6 reports the estimation results. Consistent with theoretical predictions, we find strong evidence that reputational concerns shape evaluator behavior. When an average evaluator is randomly assigned to a committee where peers are, on average, one standard deviation better than expected, they produce reports that are approximately 20 percent of a standard deviation longer. This sizable and statistically significant effect supports the view that the presence of higher-quality peers increases the effort evaluators invest in their

reports.

This mechanism offers a potential explanation for our broader finding that committees composed of more productive researchers tend to make better decisions. If the presence of top researchers induces other evaluators to exert more effort, driven by reputational concerns, then committee composition may also shape collective performance through its effect on individual behavior. In this sense, peer effects and effort complementarities, as suggested by Swank and Visser (2023), may be a key channel through which evaluator quality influences decision quality.

We also examine whether the peer effect varies with the evaluators' own research productivity. For report length, the interaction effect between evaluator productivity and committee quality is close to zero and statistically insignificant. However, when we use linguistic complexity as an alternative proxy for effort, a different pattern emerges. More productive evaluators respond more strongly to high-quality peers, producing 15 percent of a standard deviation more complex reports. While the estimated peer effect the average researcher is positive, it is not statistically distinguishable from zero.

This heterogeneity in response has important implications for the literature. If there are complementarities in evaluation effort, as some theoretical models suggest,then committees composed of multiple high-quality evaluators may generate stronger incentives to invest effort, ultimately leading to higher-quality assessments. In this view, evaluator quality not only matters at the individual level but can also enhance overall committee performance through strategic interactions and peer effects.

Alternatively, the fact that more productive evaluators increase effort in response to high-quality peers suggests that reputational incentives may be particularly salient for top researchers. Their professional standing may be more visible or subject to closer scrutiny, giving them stronger incentives to signal effort when working alongside similarly accomplished colleagues.

In contrast to the descriptive analysis, our causal estimates do not show that better-published researchers write shorter or less complex reports. This discrepancy may arise for two reasons. First, the negative association observed in the descriptive analysis may reflect aggregation bias. If other evaluators systematically increase their effort when assigned to

committees with more productive colleagues, this behavioral response could be mechanically attributed to the top researchers themselves in a pooled specification, generating a downward bias in the estimated effect of evaluator quality. Second, more productive researchers may, in expectation, be assigned to peers with lower publication records. Given the structure of the draw – whereby at most one researcher may be selected per institution – and the tendency for researchers of similar quality to cluster within institutions, evaluators from high-productivity environments are more likely to be paired with lower-productivity peers. This peer composition may suppress their effort relative to what is observed in the descriptive data.

5 Conclusion

In this paper, we provide causal evidence on how evaluator quality shapes the effectiveness of academic promotion systems. Exploiting the randomized assignment of evaluators in Italy's *Abilitazione Scientifica Nazionale*, we show that committees composed of better-published researchers apply systematically stricter standards and place greater emphasis on publication quality rather than quantity. Candidates selected by better-published committees produce more high-impact research and advance further in their careers. We argue that the presence of high-quality evaluators mainly affects outcomes by influencing committee dynamics, setting higher expectations, and prompting their peers to exert more effort in the evaluation process.

These results have important implications for the design of academic evaluation systems. First, the composition of committees materially shapes not only the stringency of assessments but also the weighting of different dimensions of candidate profiles. Moreover, improving the overall quality of evaluators may shift incentives for early-career researchers toward producing higher-impact work rather than maximizing publication counts.

However, these benefits entail substantial costs that are internalized by the evaluators. Service on Italian evaluation committees imposes a persistent decline in research output, equivalent to approximately 20–35% of annual publications. Experiencing these opportunity costs reduces subsequent participation, especially among better-published researchers, creating a tension between improving evaluation quality and sustaining the willingness of

qualified individuals to serve. While more productive evaluators attempt to optimize effort by preparing shorter and less complex reports, these adjustments are insufficient to offset the productivity losses. Over time, the selective attrition of high-quality evaluators risks diminishing both the expertise applied to assessments and the incentives for peers to maintain rigorous standards.

Our findings imply that institutional reforms aimed at raising evaluation quality must also address the costs imposed on evaluators. Policymakers should consider measures to mitigate these burdens, such as targeted teaching reductions, compensation for foregone research time, or selectively assigning complex cases to expert panels. More broadly, these findings are likely to be relevant for other domains that rely on expert peer evaluation, such as professional licensing boards in law, medicine, and engineering, where similar trade-offs between evaluation quality and the opportunity costs borne by qualified practitioners may arise. Sustaining credible, high-quality evaluation processes requires aligning incentives not only for those under review but also for those performing the evaluations.

References

Aczel, B., B. Szaszi, and A. O. Holcombe (2021, November). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review 6*(1), 14.

Alexander, D., O. Gorelkina, E. Hengel, and R. S. J. Tol (2023). Gender and the time cost of peer review. Working Paper TI 2023-044/V, Tinbergen Institute Discussion Paper.

Arts, S., N. Melluso, and R. Veugelers (2025, January). Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text. *The Review of Economics and Statistics*, 1–33.

Azoulay, P., J. S. Graff Zivin, and G. Manso (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics 42*(3), 527–554.

Azoulay, P., J. S. Graff Zivin, and J. Wang (2010, May). Superstar Extinction. *The Quarterly Journal of Economics 125*(2), 549–589.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017, April). Does the gender composition of scientific committees matter? *American Economic Review 107*(4), 1207–38.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2019a, June). Connections in scientific committees and applicants' self-selection: Evidence from a natural randomized experiment. *Labour Economics 58*, 81–97.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2019b, March). A walk on the wild side: 'Predatory' journals and information asymmetries in scientific evaluations. *Research Policy 48*(2), 462–477.

Bertocchi, G., A. Gambardella, T. Jappelli, C. A. Nappi, and F. Peracchi (2015, March). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy 44*(2), 451–466.

Brogaard, J., J. Engelberg, and C. A. Parsons (2014, January). Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics 111*(1), 251–270.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics 135*(1), 269–327.

Chan, D. C. (2021, February). Influence and Information in Team Decisions: Evidence from Medical Residency. *American Economic Journal: Economic Policy 13*(1), 106–137.

Checchi, D., G. De Fraja, and S. Verzillo (2021, September). Incentives and Careers in Academia: Theory and Empirical Analysis. *The Review of Economics and Statistics 103*(4), 786–802.

De Paola, M. and V. Scoppa (2015). Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia. *Economica 82*(325), 162–188.

Heckman, J. J. and S. Moktan (2020, June). Publishing and Promotion in Economics: The Tyranny of the Top Five. *Journal of Economic Literature 58*(2), 419–470.

Hoffman, M. and C. T. Stanton (2024, August). People, Practices, and Productivity: A Review of New Advances in Personnel Economics.

Holmstrom, B. and P. Milgrom (1991, January). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization 7*(special_issue), 24–52.

Koffi, M. (2025, July). Innovative Ideas and Gender (In)equality. *American Economic Review 115*(7), 2207–2236.

Krieger, J. L., K. R. Myers, and A. D. Stern (2023, May). How Important Is Editorial Gatekeeping? Evidence from Top Biomedical Journals. *The Review of Economics and Statistics*, 1–33.

Lazear, E. P. and S. Rosen (1981, October). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy 89*(5), 841–864.

Levy, G. (2007, March). Decision Making in Committees: Transparency, Reputation, and Voting Rules. *American Economic Review 97*(1), 150–168.

Li, D. (2017, April). Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics 9*(2), 60–92.

Mukhopadhaya, K. (2003). Jury Size and the Free Rider Problem. *Journal of Law, Economics, and Organization 19*, 24.

Myers, K. and W. Y. Tham (2023, December). Money, Time, and Grant Design.

Nieddu, M. and L. Pandolfi (2022, October). The effectiveness of promotion incentives for public employees: Evidence from Italian academia. *Economic Policy 37*(112), 697–748.

Ottaviani, M. and P. Sørensen (2001, September). Information aggregation in debate: Who should speak first? *Journal of Public Economics 81*(3), 393–421.

Priem, J., H. Piwowar, and R. Orr (2022, June). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.

Seeber, M., J. Vlegels, E. Reimink, A. Marušić, and D. G. Pina (2021, July). Does reviewing experience reduce disagreement in proposals evaluation? Insights from Marie Skłodowska-Curie and COST Actions. *Research Evaluation 30*(3), 349–360.

Stephan, P., R. Veugelers, and J. Wang (2017, April). Reviewers are blinkered by bibliometrics. *Nature 544*(7651), 411–412.

Swank, O. H. and B. Visser (2023, May). Committees as active audiences: Reputation concerns and information acquisition. *Journal of Public Economics 221*, 104875.

Vernos, I. (2013, March). Quotas are questionable. *Nature 495*(7439), 39–39.

Visser, B. and O. H. Swank (2007, February). On Committees of Experts. *The Quarterly Journal of Economics 122*(1), 337–372.

Waldinger, F. (2010, August). Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany. *Journal of Political Economy 118*(4), 787–831.

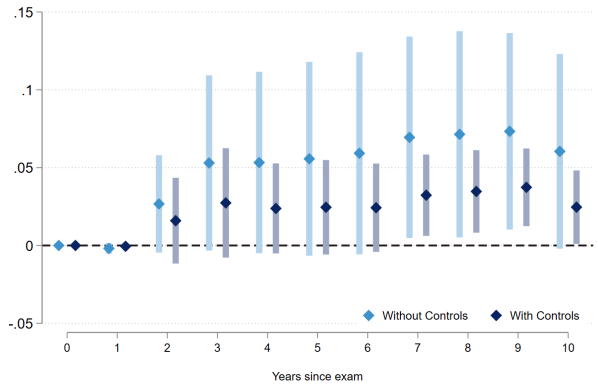
Wang, J., R. Veugelers, and P. Stephan (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy 46*(8), 1416–1436.

Zinovyeva, N. and M. Bagues (2015, April). The Role of Connections in Academic Promotions. *American Economic Journal: Applied Economics 7*(2), 264–292.

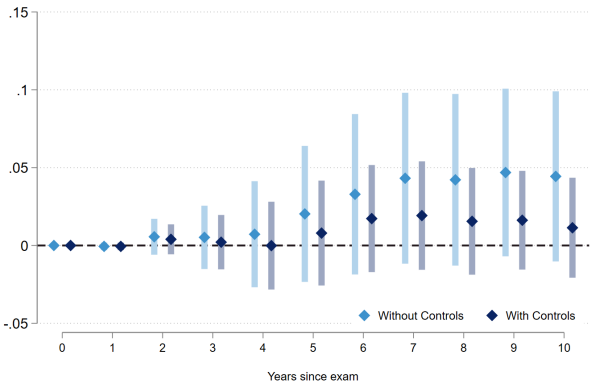
Figures

Figure 1: Impact of committee quality on the probability of future promotion

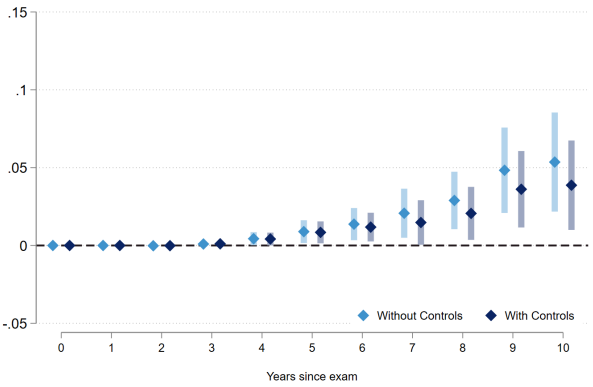
Panel A. AP Qualified Candidates, AP Promotions



Panel B. FP Qualified Candidates, FP Promotions

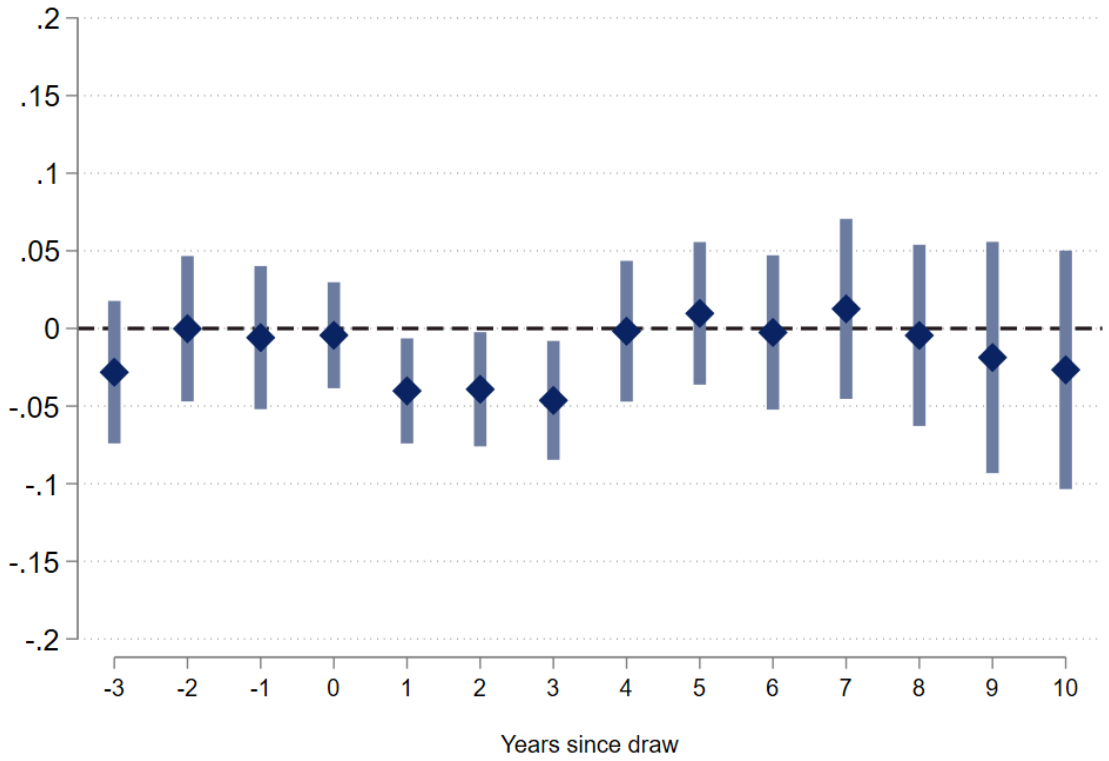


Panel C. AP Qualified Candidates, FP Promotions



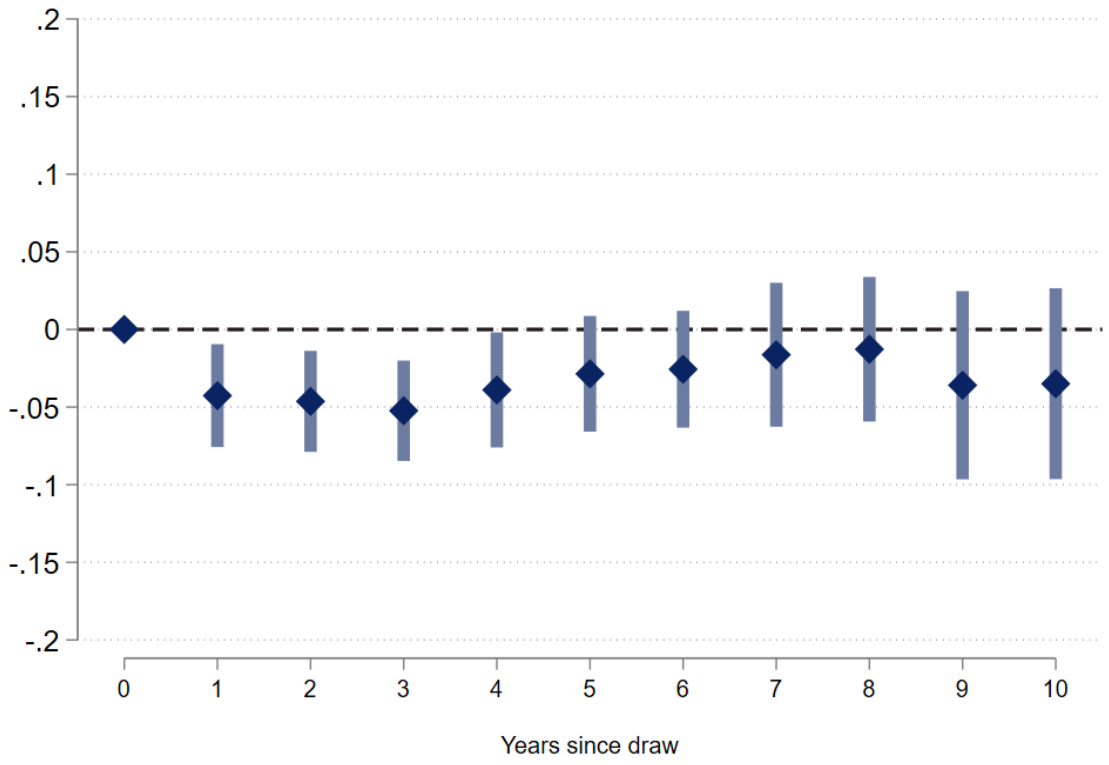
Notes: All panels use the unanimity-restricted sample for below-median-quality committees. Panel A plots the probability that an Associate Professor candidate is promoted to Associate Professor within k years; Panel B plots the probability that a Full Professor candidate is promoted to Full Professor within k years; Panel C plots the probability that an Associate Professor candidate is promoted to Full Professor within k years. Light-blue lines show estimates without controls; dark-blue lines add controls for the candidate's pre-exam affiliation, rank, and research productivity. In all figures we control for the committee's expected quality. We instrument final committee composition with the initial composition. Standard errors are clustered at the exam level.

Figure 2: Impact of sitting on the committee on total research output



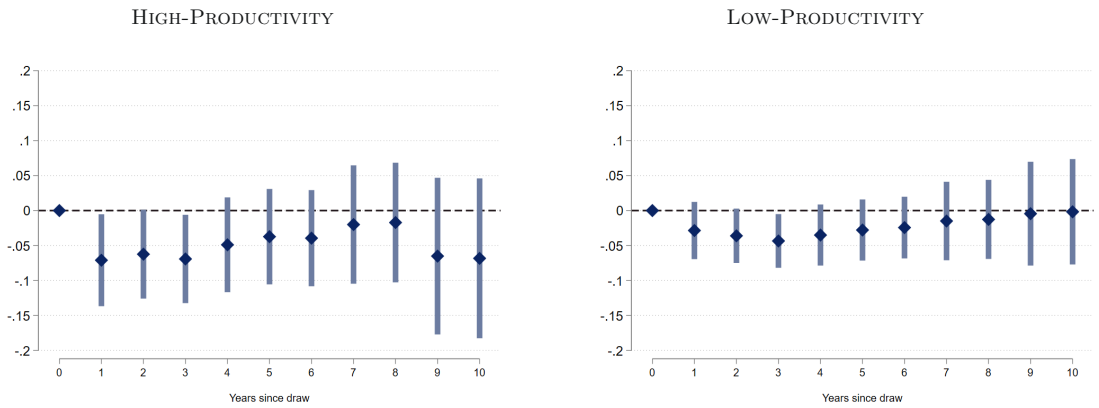
Notes: The outcome variable is total research output, normalized at the field-year level. The sample is all researchers who are observed in the pool of potential evaluators in the Italian evaluation system between 2012-2021. We control for the probability of being drawn. For years after the exam, we also control for past research production. Standard errors are clustered at the exam level.

FIGURE 3: IMPACT OF SITTING ON THE COMMITTEE ON TOTAL ACCUMULATED RESEARCH OUTPUT



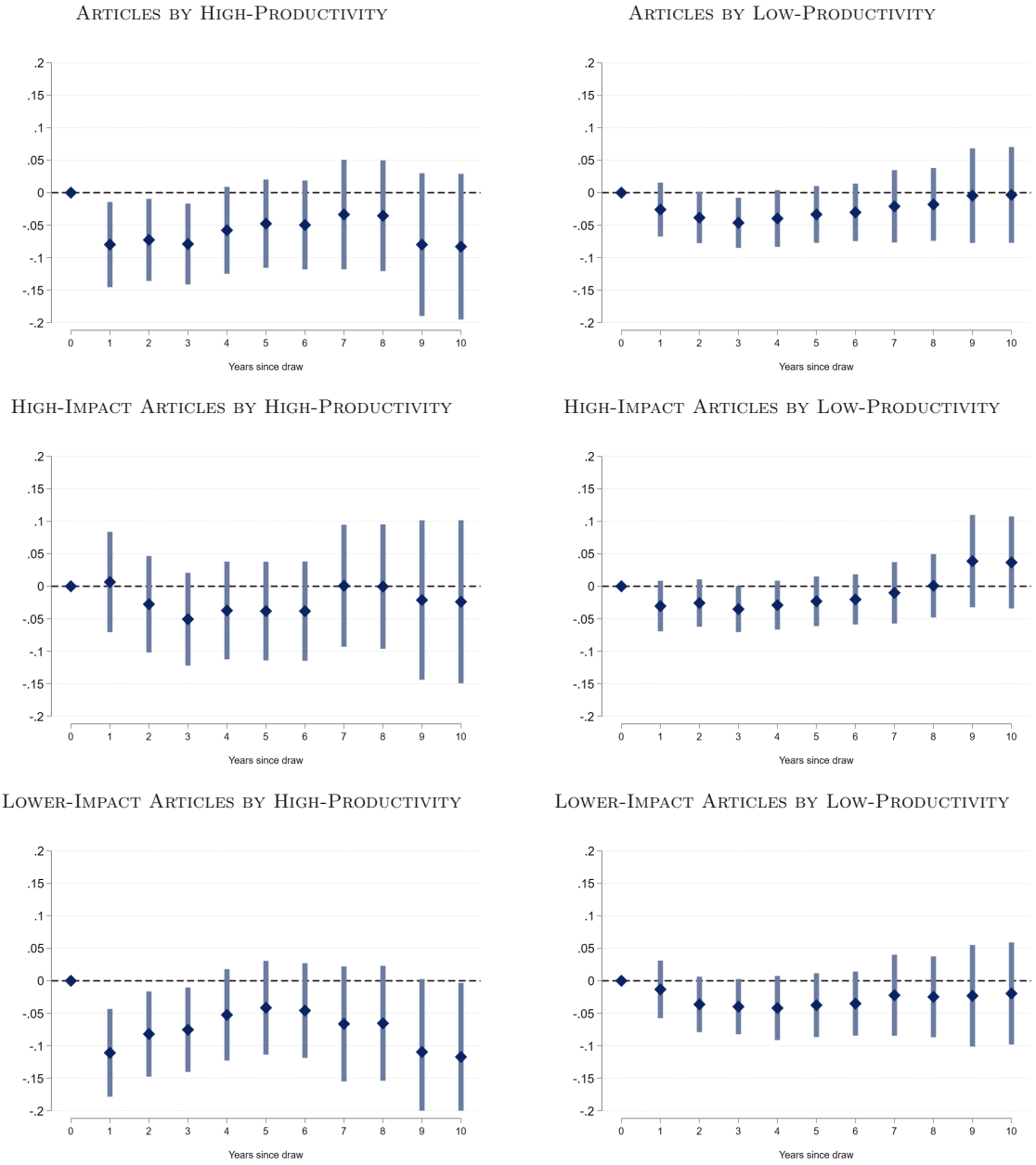
Notes: The outcome variable is accumulated total research output starting from the year of the exam, normalized at the field-year level. The sample is all researchers who are observed in the pool of potential evaluators in the Italian evaluation system between 2012-2021. We control for the probability of being drawn and past research production. Standard errors are clustered at the exam level.

Figure 4: Impact of sitting on the committee on accumulated research output by baseline research productivity



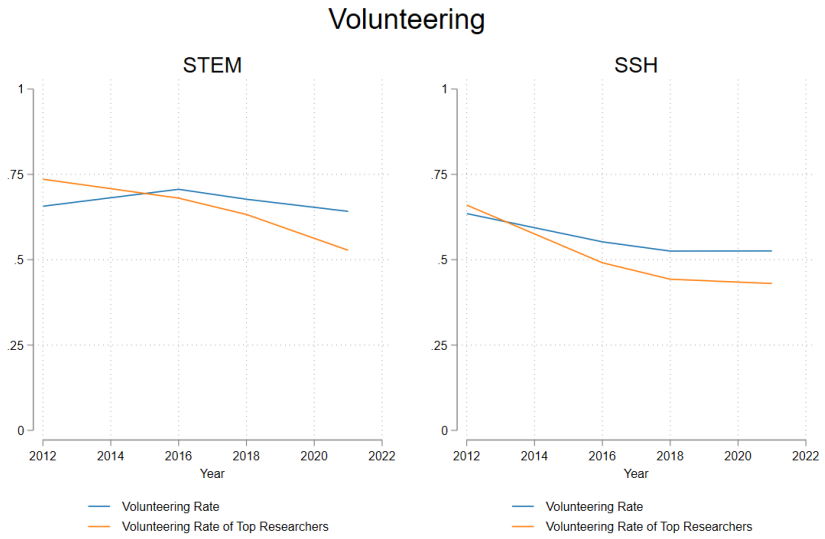
Notes: The outcome variable is accumulated total research output starting from the year of the exam, normalized at the field-year level. High-productivity researchers are defined as 75th percentile among all Italian full professors based on quality-adjusted publications in the last 10 years prior to the exam. The sample is all researchers who are observed in the pool of potential evaluators in the Italian evaluation system between 2012-2021. We control for the probability of being drawn and past research production. Standard errors are clustered at the exam level.

Figure 5: Impact of sitting on the committee on accumulated article production by productivity



Notes: The outcome variable is accumulated articles starting from the year of the exam, normalized at the field-year level. *High-impact* articles are defined by being published in journals either in the top-quartile in the Web of Science for STEM+M or in A-list journals (provided by the ministry) in SSH. We define all other articles as *lower-impact*. High-productivity researcher is defined as 75th percentile among all Italian full professors based on quality-adjusted publications in the last 10 years. The sample is all researchers who participate in the Italian evaluation system between 2012-2021. Standard errors are clustered at the exam level.

FIGURE 6: VOLUNTEERING RATE OF TOP RESEARCHERS



Notes: The figure displays trends in eligibility and volunteering rates among researchers over time. We split the sample by STEM+M and SSH. *Top researchers* are defined as professors in the 75th percentile of quality-adjusted research output within their fields among all Italian full professors over the past 10 years prior to the exam.

Tables

Table 1: Effect of committee composition on evaluation outcomes

		Overall		ASP	FP	STEM	SSH
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Candidate's Number of Publications (N_i)	0.110***	0.117***	0.107***	0.111***	0.093***	0.114***	0.096***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)
Candidate's Quality of Publications (Q_i)	0.094***	0.098***	0.098***	0.097***	0.101***	0.101***	0.091***
	(0.005)	(0.006)	(0.006)	(0.005)	(0.007)	(0.007)	(0.009)
Committee Quality (T_e)	-0.051*	-0.051*	-0.055*	-0.061*	-0.040	-0.057**	-0.045
	(0.030)	(0.030)	(0.031)	(0.032)	(0.034)	(0.029)	(0.071)
Candidate's Number of Publications x Committee Quality ($N_i \times T_e$)		-0.025**	-0.023**	-0.028**	-0.011	-0.021	-0.025
		(0.012)	(0.010)	(0.011)	(0.013)	(0.014)	(0.015)
Candidate's Number of Publications x Committee Quality ($Q_i \times T_e$)		0.032***	0.031***	0.021*	0.054***	0.021	0.055***
		(0.012)	(0.012)	(0.011)	(0.017)	(0.015)	(0.020)
Observations	69020	69020	69020	47426	21594	41395	27625
Adj. R^2	0.090	0.088	0.131	0.138	0.128	0.133	0.129
Exams	184	184	184	184	184	109	75
Candidate Connections			✓	✓	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. *Candidates' Quality of Publications* (Q_i) is the share of candidates' articles in top-quartile Web of Science journals (STEM+M) or A-list journals (SSH). *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. All candidate productivity measures are standardized within each exam, and evaluator quality is standardized within each eligibility pool before computing T_e. We instrument the final committee composition with the initial draw, before any resignations. All columns control for the expected committee quality, and columns 2-7 additionally include its interactions with candidates' publication quantity and quality. Standard errors are clustered at the exam level.

Table 2: Effect of committee composition on individual voting

	Positive Vote		
	(1)	(2)	(3)
Candidate's Number of Publications (N_i)	0.116***	0.116***	0.116***
	(0.005)	(0.005)	(0.005)
Candidate's Quality of Publications (Q_i)	0.095***	0.095***	0.095***
	(0.006)	(0.006)	(0.006)
Evaluator Quality (T_j)	-0.025**	-0.025**	-0.025**
	(0.011)	(0.011)	(0.011)
Peer Quality ($T_{e,-j}$)	-0.046**	-0.046**	-0.046**
	(0.023)	(0.023)	(0.023)
Evaluator Quality x Peer Quality			
($T_j \times T_{e,-j}$)	0.005	0.005	0.005
	(0.014)	(0.014)	(0.014)
Candidate's Number of Publications			
x Evaluator Quality ($N_i \times T_j$)		-0.004	-0.006
		(0.002)	(0.004)
Candidate's Quality of Publications			
x Evaluator Quality ($Q_i \times T_j$)		0.009***	0.008*
		(0.003)	(0.004)
Candidate's Number of Publications			
x Peer Quality ($N_i \times T_{e,-j}$)			-0.013
			(0.009)
Candidate's Quality of Publications			
x Peer Quality ($Q_i \times T_{e,-j}$)			0.028***
			(0.010)
Obs.	241669	241669	241669
Adj. R^2	0.132	0.132	0.133
Connection Controls	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the evaluator casts a positive vote. *Evaluator quality* (T_j) is the quality-adjusted publication record of evaluators over the ten years preceding the exam. Similarly, *peer quality* ($T_{e,-j}$) is the quality-adjusted publication record of all other evaluators on committee, excluding evaluator j over the ten years preceding the exam. All candidate productivity measures are standardized within each exam, and evaluator quality is standardized within each eligibility pool. We instrument the final committee composition with the initial draw, before any resignations. All columns control for the expected committee quality, candidate productivity, and connections between evaluators and candidates. Standard errors are clustered at the exam level.

Table 3: Future outcomes of selected candidates

Panel A. All qualified candidates

	Total Publications		Share High-Impact		Citations	
	(1)	(2)	(3)	(4)	(5)	(6)
Committee Quality (T_e)	0.019	0.012	0.060***	0.015	0.051*	0.039***
	(0.024)	(0.018)	(0.022)	(0.015)	(0.027)	(0.015)
Obs.	25342	25337	25342	25337	25342	25337
Adj. R^2	0.000	0.118	0.000	0.163	0.000	0.341
Pre-Exam Observables		✓		✓		✓

Panel B. Top ranked candidates

	Total Publications		Share High-Impact		Citations	
	(1)	(2)	(3)	(4)	(5)	(6)
Committee Quality (T_e)	0.007	0.006	0.052**	0.015	0.036	0.033**
	(0.024)	(0.019)	(0.023)	(0.016)	(0.028)	(0.015)
Obs.	23683	23678	23683	23678	23683	23678
Adj. R^2	-0.000	0.117	0.000	0.163	0.000	0.343
Pre-Exam Observables		✓		✓		✓

Notes: The outcome measures are accumulated for 10 years from the year of the exam, and normalized at the field level. Citations are measured between 2012-2022 on all corpus of the candidates' works. High-impact articles are defined as top-quartile publications in the Web of Science in STEM+M, and A-list articles (as defined by the Ministry) in SSH. Panel A is estimated on the sample of all qualified candidates. In Panel B we restrict the sample to include candidates qualified unanimously by committees whose below the median. The *pre-exam observables* include the candidates' affiliation, academic rank, number of publications, share of high-impact articles, and citations. The quality of the committee, T_e, is measured by the average quality-adjusted publications of evaluators over the past 10 years prior to the exam. The final composition of the committee is instrumented by the initial committee composition, due to potentially endogenous resignations. We control for the expected quality of the committee. Standard errors are clustered at the exam level.

Table 4: Probability of volunteering again

	Volunteer Again		
Drawn_j	-0.190***	-0.190***	-0.197***
	(0.013)	(0.013)	(0.013)
$\text{Drawn}_j \times$ Evaluator Quality			
($\text{Drawn}_j \times \text{T}_j$)			-0.021*
			(0.012)
Observations	14536	14536	13652
R^2	0.016	0.015	0.016
Exam FE		✓	✓

Notes: The outcome variable is a dummy variable indicating whether a researcher who was observed in the pool of potential evaluators is again observed volunteering in subsequent years. The productivity of researchers, T_j, is measured by the quality-adjusted publications of evaluators over the past 10 years prior to the exam. The final draw status of committee members is instrumented by the initial random draw. We control for the probability that a researcher gets drawn, and the interaction between the draw probability and the productivity of researchers. Standard errors are clustered at the exam level.

Table 5: Textual features of evaluation reports and characteristics of candidates and evaluators

	Words		Average IDF	
	(1)	(2)	(3)	(4)
Evaluator Quality (T_j)	-0.010	-0.053**	-0.036	-0.064***
	(0.030)	(0.025)	(0.028)	(0.023)
Candidate's Number of Publications (N_i)	0.057***		0.017*	
	(0.009)		(0.009)	
Candidate's Quality of Publications (Q_i)	0.038***		0.029***	
	(0.010)		(0.007)	
Coauthors_{ij}	0.127**	0.081***	0.148***	0.110***
	(0.057)	(0.026)	(0.041)	(0.030)
$\text{Same Affiliation}_{ij}$	0.110***	0.091***	0.051	0.066**
	(0.039)	(0.035)	(0.036)	(0.032)
Same Field_{ij}	0.175***	0.173***	0.069**	0.231***
	(0.035)	(0.047)	(0.031)	(0.043)
Obs.	241744	241744	241744	241744
R^2	0.013	0.581	0.004	0.519
Candidate FE		✓		✓

Notes: The table shows that better-published researchers write shorter and less complex reports. The observations are at the evaluation level (candidate-evaluator level). All measures are normalized at the field-category level. There are 184 academic fields and 2 application categories: Associate and Full. T_j is the quality-adjusted research output of evaluators in the past 10 years. N_i is the candidate's number of publications, Q_i is the candidate's share of high-impact articles. High-impact articles are defined as top-quartile publications in the Web of Science in STEM+M, and A-list articles (provided by the Ministry) in SSH. Standard errors are clustered at the exam level.

Table 6: Impact of Committee Composition on Evaluator Report Writing

	Words		Average IDF	
	(1)	(2)	(3)	(4)
Evaluator Quality (T_j)	-0.017	-0.056	-0.014	-0.026
	(0.040)	(0.042)	(0.035)	(0.042)
Peer Quality ($T_{e,-j}$)	0.211**	0.201**	0.125	0.130
	(0.095)	(0.095)	(0.082)	(0.087)
Evaluator Quality x Peer Quality				
($T_j \times T_{e,-j}$)		0.058		0.150**
		(0.071)		(0.073)
Obs.	241744	241744	241744	241744
R^2	0.017	0.020	0.006	0.005
Connection Controls	✓	✓	✓	✓
Candidate Controls	✓	✓	✓	✓

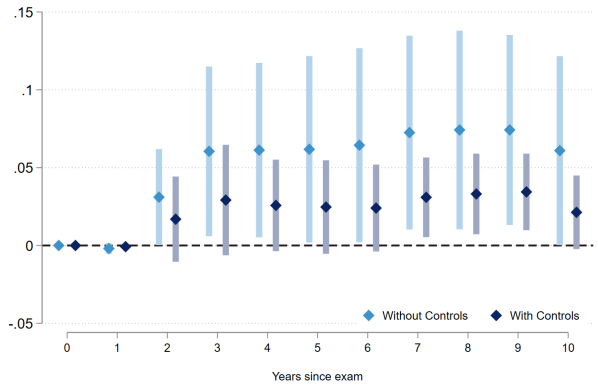
Notes: The observations are at the evaluation level (candidate-evaluator level). All measures are normalized at the macro field-category level. Note that this is different from the results in Table 5. Here it no longer makes sense to normalize at the field-category level since we are exploiting a shock that is highly correlated at the exam-level. There are 184 fields and 2 categories: Associate and Full. $T_{e,-j}$ is the average quality-adjusted research output of other evaluators on the committee in the past 10 years. Standard errors are clustered at the exam level.

Appendix

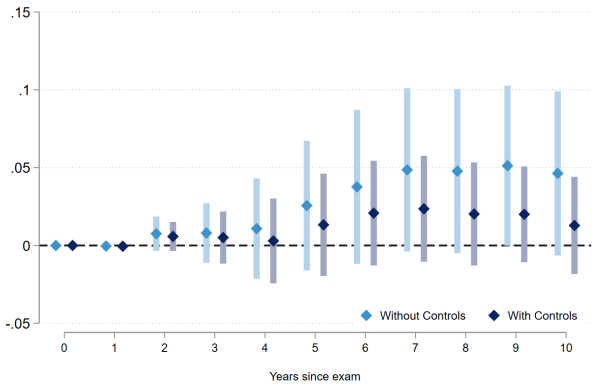
Appendix Figures

Figure A1: Impact of committee quality on the probability of future promotion

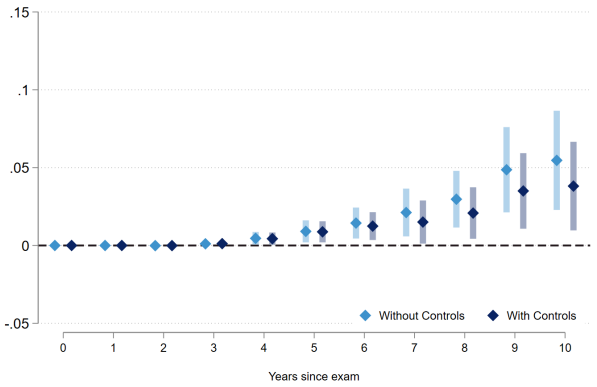
Panel A. AP Qualified Candidates, AP Promotions



Panel B. FP Qualified Candidates, FP Promotions

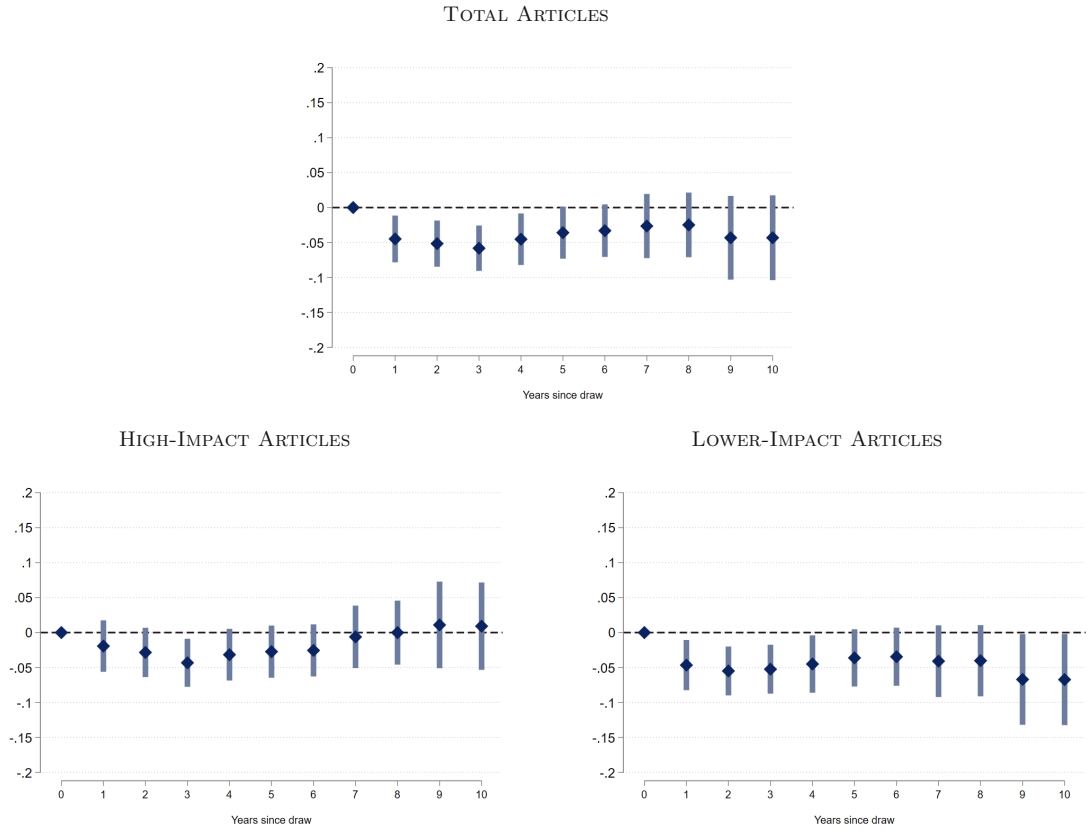


Panel C. AP Qualified Candidates, FP Promotions



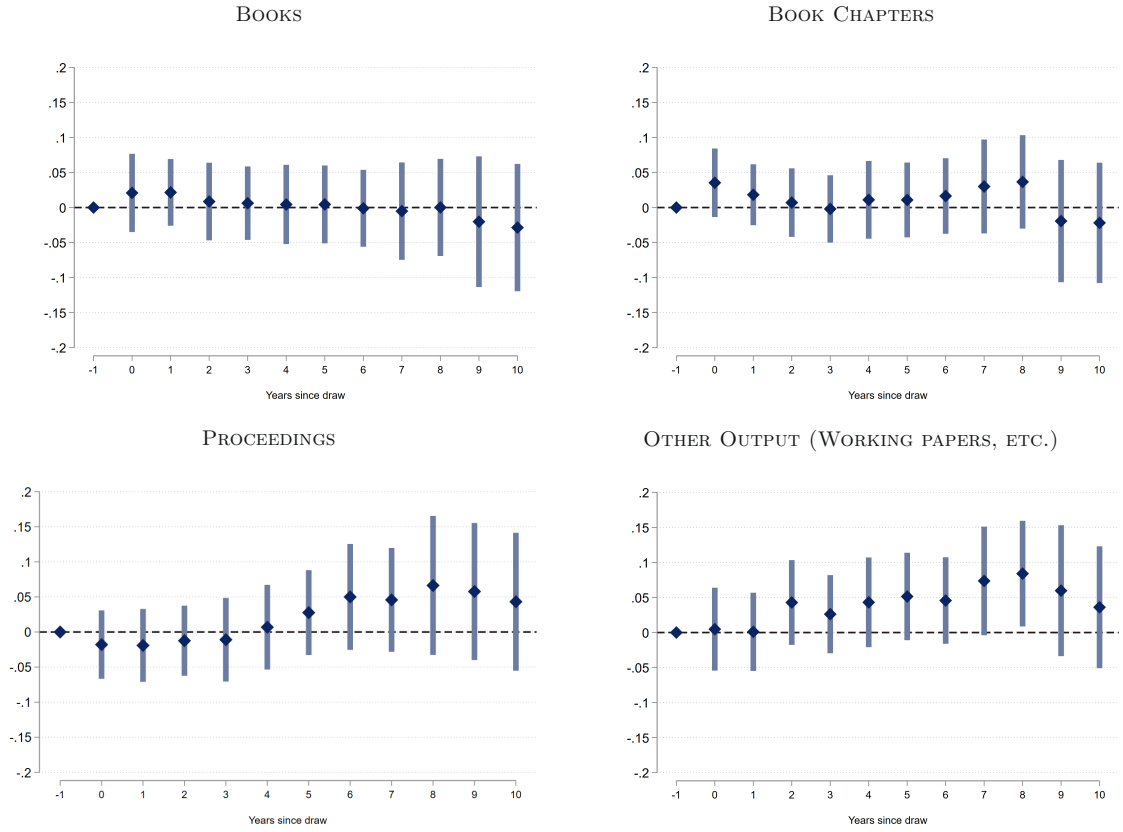
Notes: All panels use the sample of all qualified candidates. Panel A plots the probability that an Associate Professor candidate is promoted to Associate Professor within k years; Panel B plots the probability that a Full Professor candidate is promoted to Full Professor within k years; Panel C plots the probability that an Associate Professor candidate is promoted to Full Professor within k years. Light-blue lines show estimates without controls; dark-blue lines add controls for *pre-exam observables* including the candidates' affiliation, academic rank, number of publications, share of high-impact articles, and citations. In all figures we control for the committee's expected quality. We instrument final committee composition with the initial composition. Standard errors are clustered at the exam level.

Figure A2: Impact of sitting on the committee on accumulated article production



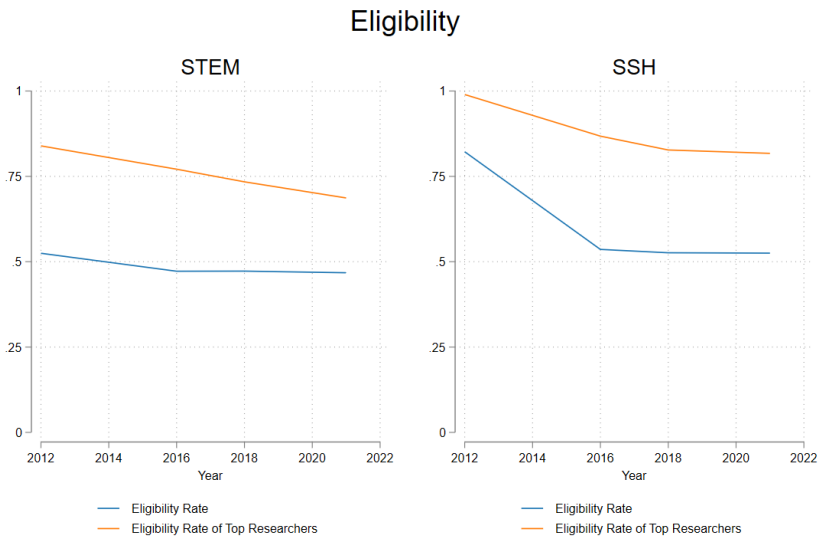
Notes: The outcome variable is accumulated articles starting from the year of the exam, normalized at the field-year level. *High-impact* articles are defined by being published in journals either in the top-quartile in the Web of Science for STEM+M or in A-list journals (provided by the ministry) in SSH. We define all other articles as *lower-impact*. The sample is all researchers who participate in the Italian evaluation system between 2012-2021. Standard errors are clustered at the exam level.

Figure A3: Impact of sitting on the committee on accumulated other research production



Notes: The figure shows the estimates of the impact of being randomly drawn to sit on the committee on future research output. The measures of productivity are normalized at the field level. The sample is all researchers who participate in the Italian evaluation system between 2012-2021.

FIGURE A4: ELIGIBILITY OF TOP RESEARCHERS



Notes: The figure displays trends in eligibility and volunteering rates among researchers over time. We split the sample by STEM+M and SSH. *Top researchers* are defined as professors in the 75th percentile of quality-adjusted research output within their fields among all Italian full professors over the past 10 years prior to the exam.

Appendix Tables

Table A1: Draw procedure and potential evaluators

Panel A. Pool of Potential Evaluators

	Mean	SD	Min	Max	P10	Median	P90
Potential Evaluators	26.54	16.70	10	138	12	21	48
Female Potential Evaluators	5.82	5.09	0	35	1	4	13
Initial Draw	5.00	0.00	5	5	5	5	5
Resignation	0.45	0.72	0	5	0	0	1
Final Draw	4.97	0.20	4	6	5	5	5
Participating Universities	16.03	7.00	5	53	9	14	25
Number of Exams	939						

Panel B. Research Profiles of Potential Evaluators

	Mean	SD	Min	Max	P10	Median	P90
Exam Participations	1.68	1.00	1	7	1	1	3
On Initial Committee	0.22	0.37	0	1	0	0	1
Resigned	0.02	0.12	0	1	0	0	0
On Final Committee	0.21	0.36	0	1	0	0	1
Publications Matched	0.94	0.24	0	1	1	1	1
Accumulated Works	139.51	164.18	0	633	14	81	348
Accumulated Articles	136.31	161.56	0	621	13	77	344
Accumulated High-Quality Articles	57.94	82.75	0	317	3	25	156
Number of Researchers	14886						

Notes: The sample is all Professors who volunteered to participate as an evaluator in Abilitazione between 2012-2023. In the 2012 wave one out of five evaluators was an external researcher, affiliated with a non-Italian University. *Participating Universities* counts the number of unique affiliations of researchers who participate in the exam.

Table A2: Candidates in the 2012 exams

Panel A. All candidates

	Mean	SD	Min	Max	P10	Median	P90
Female Candidate	0.38	0.49	0	1	0	0	1
Italian Affiliation	0.67	0.47	0	1	0	1	1
Tenured Position Prior to Application	0.53	0.50	0	1	0	1	1
First Year of Publication	1996.53	7.50	1976	2009	1986	1998	2005
Accumulated Works	62.89	58.29	1	334	13	45	133
Accumulated Articles	35.46	40.01	0	226	4	21	83
Accumulated High-Quality Articles	13.64	19.77	0	114	0	6	35
Number of Applications	69020						
Number of Candidates	46329						

Panel B. Candidates to Associate

	Mean	SD	Min	Max	P10	Median	P90
Female Candidate	0.41	0.49	0	1	0	0	1
Italian Affiliation	0.62	0.49	0	1	0	1	1
Tenured Position Prior to Application	0.43	0.50	0	1	0	0	1
First Year of Publication	1998.43	6.66	1976	2009	1989	2000	2006
Accumulated Works	52.02	48.80	1	334	12	38	107
Accumulated Articles	28.62	32.72	0	226	4	18	64
Accumulated High-Quality Articles	10.89	16.31	0	114	0	5	28
Promoted to ASP within 10 years	0.46	0.50	0	1	0	0	1
Promoted to FP within 10 years	0.09	0.28	0	1	0	0	0
Number of Applications	47426						
Number of Candidates	34643						

Panel C. Candidates to Full

	Mean	SD	Min	Max	P10	Median	P90
Female Candidate	0.31	0.46	0	1	0	0	1
Italian Affiliation	0.77	0.42	0	1	0	1	1
Tenured Position Prior to Application	0.73	0.45	0	1	0	1	1
First Year of Publication	1992.37	7.57	1976	2009	1982	1993	2002
Accumulated Works	86.75	69.30	1	334	20	68	176
Accumulated Articles	50.49	49.36	0	226	7	33	118
Accumulated High-Quality Articles	19.70	24.74	0	114	0	10	52
Promoted to FP within 10 years	0.36	0.48	0	1	0	0	1
Number of Applications	21594						
Number of Candidates	16747						

Notes: The sample is all Professors who volunteered to participate as an evaluator in Abilitazione between 2012-2023. In the 2012 wave one out of five evaluators was an external researcher, affiliated with a non-Italian University. *Participating Universities* counts the number of unique affiliations of researchers who participate in the exam.

Table A3: Eligibility

	Eligible		Not Eligible		Difference	
	Mean	SD	Mean	SD	Diff	T-stat
Female	0.22	0.42	0.24	0.43	0.02***	4.97
Total Works	1.07	1.40	0.25	1.00	-0.82***	-77.31
High-Impact Publications	1.02	1.51	0.11	0.98	-0.92***	-83.35
Articles	1.06	1.40	0.25	1.00	-0.81***	-77.26
Article in WOS	0.78	1.41	0.20	1.05	-0.58***	-53.88
Total AIS	0.68	1.46	0.13	1.03	-0.55***	-49.91
Low-Impact Publications	0.85	1.39	0.30	1.06	-0.55***	-51.43
Predatory Publications	0.04	1.03	0.04	1.05	0.00	0.10
Books	0.22	1.34	0.05	1.13	-0.16***	-15.01
Chapters	0.32	1.43	0.09	1.08	-0.23***	-20.71
Other Works	0.13	1.26	0.07	1.10	-0.06***	-5.94
Total Works (Past 10 Years)	0.79	1.34	-0.02	0.97	-0.82***	-80.36
High-Impact Publications (Past 10 Years)	0.79	1.42	-0.09	0.93	-0.88***	-84.93
Articles (Past 10 Years)	0.79	1.34	-0.02	0.97	-0.82***	-80.43
Article in WOS (Past 10 Years)	0.59	1.34	0.02	1.00	-0.57***	-56.00
Total AIS (Past 10 Years)	0.52	1.37	-0.01	0.97	-0.53***	-51.57
Low-Impact Publications (Past 10 Years)	0.60	1.33	0.04	1.00	-0.57***	-55.50
Predatory Publications (Past 10 Years)	0.03	1.00	0.03	1.03	0.00	0.35
Books (Past 10 Years)	0.17	1.28	0.01	1.05	-0.15***	-14.82
Chapters (Past 10 Years)	0.21	1.32	0.01	1.01	-0.20***	-19.38
Other Works (Past 10 Years)	0.09	1.15	0.03	1.05	-0.05***	-5.13
Observations	28205		23939		52144	

Notes: Evaluator eligibility is tied to productivity prior to the exam. All measures are normalized at the field-year level, on the sample of all Italian full professors.

Table A4: Volunteering conditional on being eligible

	Volunteered		Not Volunteered		Difference	
	Mean	SD	Mean	SD	Diff	T-stat
Female	0.22	0.41	0.23	0.42	0.01*	2.29
Total Works	1.21	1.42	0.96	1.37	-0.25***	-14.96
High-Impact Publications	1.14	1.55	0.93	1.47	-0.21***	-11.29
Articles	1.21	1.42	0.95	1.36	-0.25***	-15.08
Article in WOS	0.85	1.44	0.73	1.38	-0.12***	-7.08
Total AIS	0.69	1.43	0.67	1.48	-0.03	-1.57
Low-Impact Publications	0.98	1.42	0.75	1.35	-0.23***	-13.85
Predatory Publications	0.09	1.10	0.00	0.98	-0.08***	-6.52
Books	0.26	1.46	0.18	1.25	-0.08***	-4.84
Chapters	0.37	1.50	0.29	1.38	-0.08***	-4.85
Other Works	0.15	1.31	0.11	1.22	-0.04*	-2.40
Total Works (Past 10 Years)	0.96	1.36	0.66	1.32	-0.30***	-18.67
High-Impact Publications (Past 10 Years)	0.93	1.45	0.68	1.39	-0.25***	-14.65
Articles (Past 10 Years)	0.96	1.36	0.66	1.32	-0.30***	-18.73
Article in WOS (Past 10 Years)	0.69	1.36	0.52	1.32	-0.17***	-10.42
Total AIS (Past 10 Years)	0.56	1.33	0.49	1.41	-0.07***	-4.17
Low-Impact Publications (Past 10 Years)	0.75	1.37	0.49	1.29	-0.27***	-16.74
Predatory Publications (Past 10 Years)	0.07	1.05	-0.00	0.96	-0.08***	-6.18
Books (Past 10 Years)	0.21	1.38	0.13	1.19	-0.08***	-5.01
Chapters (Past 10 Years)	0.25	1.38	0.17	1.28	-0.08***	-4.93
Other Works (Past 10 Years)	0.10	1.18	0.07	1.12	-0.03	-1.84
Observations	12279		15904		28183	

Notes: The table shows that better-published researchers are more likely to volunteer, conditional on eligibility. All measures are normalized at the field-year level, on the sample of all Italian full professors.

Table A5: Resignations

	Resigned		Not Resigned		Difference	
	Mean	SD	Mean	SD	Diff	T-stat
Female	0.21	0.41	0.25	0.43	0.04	1.89
Total Works	1.00	1.39	0.93	1.36	-0.07	-1.00
High-Impact Publications	1.03	1.64	0.84	1.45	-0.19*	-2.26
Articles	0.99	1.39	0.92	1.36	-0.07	-0.97
Article in WOS	0.67	1.26	0.73	1.36	0.06	0.85
Total AIS	0.60	1.45	0.58	1.33	-0.02	-0.28
Low-Impact Publications	0.74	1.24	0.79	1.36	0.05	0.74
Predatory Publications	0.20	1.38	0.13	1.19	-0.08	-1.04
Books	0.21	1.20	0.19	1.26	-0.02	-0.37
Chapters	0.44	1.59	0.32	1.42	-0.12	-1.49
Other Works	0.26	1.47	0.20	1.44	-0.06	-0.80
Total Works (Past 10 Years)	0.79	1.26	0.72	1.32	-0.07	-1.00
High-Impact Publications (Past 10 Years)	0.79	1.41	0.67	1.38	-0.12	-1.64
Articles (Past 10 Years)	0.78	1.27	0.72	1.32	-0.06	-0.92
Article in WOS (Past 10 Years)	0.55	1.20	0.58	1.30	0.02	0.35
Total AIS (Past 10 Years)	0.49	1.30	0.46	1.27	-0.03	-0.44
Low-Impact Publications (Past 10 Years)	0.60	1.20	0.59	1.30	-0.01	-0.10
Predatory Publications (Past 10 Years)	0.14	1.24	0.10	1.14	-0.04	-0.67
Books (Past 10 Years)	0.21	1.32	0.15	1.26	-0.05	-0.77
Chapters (Past 10 Years)	0.36	1.49	0.22	1.32	-0.14	-1.77
Other Works (Past 10 Years)	0.17	1.25	0.16	1.35	-0.01	-0.09
Observations	406		4125		4531	

Notes: The table shows that there is little correlation between observable researcher characteristics and resignations.

Table A6: The impact of committee quality on application withdrawal

	Application Withdrawn		
N_i	-0.034***	-0.034***	-0.034***
	(0.002)	(0.002)	(0.002)
Q_i	-0.026***	-0.026***	-0.026***
	(0.003)	(0.003)	(0.003)
$Predatory_i$	0.040***	0.041***	0.025***
	(0.011)	(0.010)	(0.009)
T_e	0.019	0.019	
	(0.015)	(0.016)	
$T_e \times N_i$	-0.006	-0.006	-0.006
	(0.007)	(0.006)	(0.006)
$T_e \times Q_i$	0.002	0.002	0.002
	(0.010)	(0.010)	(0.010)
$T_e \times Predatory_i$		-0.003	0.018
		(0.024)	(0.022)
Observations	69020	69020	69020
R^2	0.014	0.014	0.014
Exam FE			✓

Notes: The table shows how publications of prospective candidates in Beall's list journals are effected by the quality of committee in the corresponding field, by year since the initial assignment of committee. The sample includes candidates who initially submitted their application in 2012 but withdrew it after committee was formed and candidates in 2012 who were accessed by the committee but re-applied in 2013. The quality of the committee, T_e, is measured by the average number of quality-adjusted publications of evaluators over the past 10 years prior to the exam. The final composition of the committee is instrumented by the initial committee composition, before resignations took place. We control for the expected quality of the committee. Standard errors are clustered at the field level.

A Conceptual framework

In this section, we present a conceptual framework for our empirical analysis, which helps clarify how evaluators balance their evaluation tasks against research tasks. We draw insights from the multitasking literature in personnel economics (Lazear and Rosen, 1981; Holmstrom and Milgrom, 1991), which emphasizes that agents allocate effort across competing tasks based on their preferences, productivity, and opportunity costs. The conceptual framework highlights conditions under which higher-ability evaluators may spend more or less effort on evaluations, and illustrates how this choice affects the quality of their evaluations and the amount of research output they forgo.

A.1 The setup

Consider an evaluator tasked with assessing candidates for promotion.[16] The evaluator must exert costly effort to gather information about candidates' quality, where this cost is determined by the opportunity cost of time in terms of the evaluator's research output.

Let the performance of evaluator in the evaluation task V_j depend positively on the time she spends evaluating t_j and her ability a_j:

$$V_j = V(t_j, a_j),$$

with $V_1 > 0$, $V_2 > 0$, and $V_{11} < 0$, where numerical subscripts denote partial derivatives with respect to the corresponding arguments.

Research output depends positively on the time devoted to research $(T - t_j)$ and ability:

$$R_j = R(T - t_j, a_j),$$

where T is the total time endowment and $R_1 > 0$, $R_2 > 0$, and $R_{11} < 0$. Thus evaluation effort t_j directly reduces research production by decreasing available research time.

The evaluator derives utility from both evaluation quality and research output. Her

16. In what follows, we arbitrary refer to the evaluator as *she/her/hers*.

utility function is:

$$U_j = \beta(a_j)V(t_j, a_j) + \gamma(a_j)R(T - t_j, a_j),$$

where $\beta(a_j) > 0$ and $\gamma(a_j) > 0$ represent weights placed on evaluation quality and research output, respectively. The individual's ability can influence these weights, potentially aligning or misaligning her incentives with those of the social planner.[17]

A.2 Optimal choices and empirical implications

Given the evaluation task, the evaluator chooses t_j to maximize utility, satisfying the first-order condition:

$$\frac{\partial U_j}{\partial t_j} = \beta(a_j)V_1 - \gamma(a_j)R_1 = 0.$$

The comparative static of this decision with respect to ability a_j is:

$$\frac{dt_j^*}{da_j} = -\frac{\beta_1 V_1 - \gamma_1 R_1 + \beta V_{12} - \gamma R_{12}}{\beta V_{11} + \gamma R_{11}}. \tag{7}$$

Since $\beta V_{11} + \gamma R_{11} < 0$, higher-ability evaluators spend more time evaluating only if:

$$\beta_1 V_1 - \gamma_1 R_1 + \beta V_{12} - \gamma R_{12} > 0. \tag{8}$$

Thus, a positive relationship between evaluation effort and ability implies that more able researchers either place a greater importance on their contribution to evaluation than to research ($\beta_1 > \gamma_1$) or have a comparative advantage in evaluation task ($V_{12} > R_{12}$). In contrast, a negative relationship suggests their higher opportunity costs ($R_{12} > V_{12}$) or weaker intrinsic motivation for evaluation ($\gamma_1 > \beta_1$).

The sensitivity of evaluation performance to the ability of the evaluator takes the follow-

17. While high-ability researchers may have stronger intrinsic motivation to uphold rigorous evaluation standards, the greater observability and more direct rewards of research output can lead them to place relatively more weight on research than evaluation.

ing form:

$$\frac{\partial V(t_j^*, a_j)}{\partial a_j} = V_1 \frac{dt_j^*}{da_j} + V_2 = 0. \tag{9}$$

Equation 9 suggests that even if high-ability individuals allocate relatively less time to evaluations, they may still provide better evaluations if their absolute advantage in performing the evaluation task V_2 is large enough.

We can also assess how the impact of evaluation task on research productivity depends on ability:

$$\frac{\partial[R(T, a_j) - R(T - t_j^*, a_j)]}{\partial a_j} = R_1 \frac{dt_j^*}{da_j} + [R_2(T, a_j) - R_2(T - t_j^*, a_j)] = R_1 \frac{dt_j^*}{da_j} + R_{12} t_j^*. \tag{10}$$

Expression (10) suggests that the opportunity cost parameter $R_{12} > 0$ affects the productivity-loss gradient in two opposing ways. First, a higher $R_{12} > 0$ directly increases the marginal loss per unit of evaluation time for higher-ability evaluators. Second, higher opportunity costs reduce evaluation time, lowering the loss of research productivity. The net effect on the productivity-loss gradient is thus theoretically ambiguous. This implies that finding no significant relationship between ability and loss of research productivity in empirical analysis does not necessarily indicate equal opportunity costs among evaluators; it may reflect offsetting direct and indirect effects.

In sum, it is theoretically possible for higher-ability researchers to produce better evaluations, forgo less research output, and spend less time on evaluations. In such cases, the social planner may strictly prefer to assign the most capable individuals to evaluation tasks. However, this outcome is not guaranteed. Despite their greater skill and possibly stronger intrinsic motivation, the high opportunity cost of their time may discourage them from investing sufficient effort – or even from participating at all. In this case, they may allocate less time to evaluations, deliver lower-quality evaluations despite their higher ability, or opt out of evaluation tasks entirely.

In our empirical analysis, we proxy ability with bibliometric indicators based on evaluators' previous publication record. We then first test whether better-published researchers deliver systematically better evaluations, signaling higher intrinsic motivation ($\beta_1 > 0$) or

comparative advantage ($V_{12} > 0$). Second, we analyze whether better-published evaluators forego relatively more or less research output, indicating either higher opportunity costs ($R_{12} > 0$) or effective time optimization ($\frac{dt_j^*}{da_j} < 0$). Third, we explore whether there is evidence that the time evaluators spend on evaluation task varies with the measures of their research productivity.

Our theoretical framework abstracts from several real-world complexities. First, evaluation time might also distract researchers from other valuable activities, such as mentoring junior researchers or collaborating on coauthored work (Azoulay et al., 2010; Waldinger, 2010). This may potentially affect not only evaluators' own research output but also the productivity of other scholars. Second, the model omits committee-level strategic interactions, in which the effort choice of the evaluator may respond to expectations about peers' contributions. Such dynamics could emerge in response to reputational concerns or potential for free-riding (Mukhopadhaya, 2003; Visser and Swank, 2007; Swank and Visser, 2023). In the empirical analysis, we explore both possibilities.

B Robustness Checks

In this section, we describe other dimensions of candidate research profiles, and perform a series of robustness checks to make sure our results are not being driven by decisions in how we construct our measures.

B.1 Candidate research profiles

In the main analysis, we characterize candidate research output along two core dimensions: *quality* and *quantity*. While research profiles are inherently multidimensional, these two measures capture distinct aspects of research production that are both salient for evaluation and relevant for policy. Moreover, they are only moderately correlated, ensuring that they provide complementary rather than redundant information. Quantity reflects the scale of a candidate's scholarly output, while quality aims to capture the perceived scientific contribution and impact of this output.

The restriction to these two dimensions is motivated by both conceptual clarity and empirical considerations. First, quality and quantity are central to many hiring and promotion decisions, and are explicitly targeted by research evaluation policies. Second, a range of other candidate attributes — such as seniority, the diversity of research topics, the size of collaboration networks, — are often highly correlated with these two core dimensions. As such, including them in the main analysis would add complexity without substantively altering the interpretation of results.

Quantity The total number of publications authored by the candidate prior to the exam. To account for systematic differences in modes of scholarly communication across fields, we include a broad set of outputs such as journal articles, books, book chapters, and conference proceedings.

Quality/Impact (Journal-based) A measure of research quality proxied by the share of a candidate's articles published either in the top quartile of the Web of Science journal rankings for STEM+M fields or in A-list journals designated by the ministry in Social Sciences

and Humanities. This dimension reflects the perceived impact and selectivity of publication outlets.

Quality/Impact (Article Influence Score) An alternative measure of journal quality based on the *Article Influence Score* (AIS), which captures the average influence of articles in a journal over the five years following publication. The AIS is derived from the Eigenfactor metric and adjusts for differences in citation practices across fields, providing a field-normalized indicator of journal prestige and scholarly impact.

Quality/Impact (Citation-based) A measure of the individual impact of a candidate's research, calculated as the average number of citations per publication. This metric reflects the recognition and influence of the candidate's work within the academic community.

Number of Collaborators The average number coauthors per publication, indicating the breadth of a candidate's research network and collaborative activity.

Seniority An indicator of career stage, proxied by the number of years since the first recorded publication. This captures the candidate's accumulated experience in academic research.

Topic Dispersion The number of topics in which the author has published, divided by the total number of publications. This measures how concentrated or diversified a candidate's research topics are.

Novelty An index of the extent to which a candidate's research introduces combinations of words or phrases are new in the literature (Arts et al., 2025). This captures the originality of the research profile.

Table A7 reports the correlations between the different dimensions of candidates' research profiles. Table A8 reports the correlations between qualification and each dimension. Individually, all measures are correlated with qualification in the expected direction: the marginal effects of quantity, journal quality, citations, and novelty are positive, whereas

the number of collaborators and topic dispersion are negatively correlated. The relationship between seniority and qualification follows an inverse U-shape, indicating that returns to experience diminish at later stages. Among all measures, *Quantity* and *Journal Impact* display the strongest predictive power, consistent with our rationale for focusing on these dimensions in the main analysis. In the conditional specification (column 8), the journal-based measure dominates the other quality indicators: its point estimate increases, while the coefficient on citations becomes substantially smaller, and the Article Influence Score is no longer statistically distinguishable from zero.

Table A7: Correlation between dimensions of candidate research profiles

	Quantity	Journal Impact	AIS	Citations	Number of Collaborators	Seniority	Novelty	Topic Dispersion
Quantity	1	-.134318	-.051674	.170893	.0095592	.3223924	.1993816	-.1936371
Journal Impact	-.134318	1	.4739293	.1358741	.0186125	-.1346468	.10844	-.0863268
AIS	-.051674	.4739293	1	.1678936	.0486562	-.1162266	.1009231	-.0575354
Citations	.170893	.1358741	.1678936	1	.1612483	.0886833	.6008648	-.2867082
Number of Collaborators	.0095592	.0186125	.0486562	.1612483	1	-.0124691	.1119251	-.0077701
Seniority	.3223924	-.1346468	-.1162266	.0886833	-.0124691	1	.0590485	-.0450409
Novelty	.1993816	.10844	.1009231	.6008648	.1119251	.0590485	1	-.2408683
Topic Dispersion	-.1936371	-.0863268	-.0575354	-.2867082	-.0077701	-.0450409	-.2408683	1

Notes: The sample is all candidates who applied for qualification in the 2012 wave of Abilitazione. All candidate productivity measures are standardized within each exam.

B.2 Robustness of candidate publication quantity and quality

We analyze two dimensions of candidates' research profiles: quantity and quality. We define quality as the total number of works that appear on the candidate's CV, and quality as the share of articles that we classify as high-impact. We consider an article high-impact if it is published in a journal that is in the top-quartile Web of Science journals for STEM+M fields or in a journal that is on the A-list of journals provided by the ministry in Social Science and Humanities.

We begin by confirming that our results are robust if we use an alternative data source. Column 2 of Table A9 reports the results of the same analysis using OpenAlex and IRIS publication data, rather than CVs. We observe that with this dataset we explain less variation in the outcome variable since the R^2 is lower, and there is a slightly lower return to publication quality. This is likely due to the fact that we are no longer exploiting information that was directly provided to evaluators. The key coefficient of interest that we may expect to

Table A8: Correlation between dimensions of candidate research profiles and qualification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Quantity	0.103***								0.096***
	(0.005)								(0.005)
Quality: Journal Impact		0.083***							0.091***
		(0.005)							(0.005)
Quality: Article Influence Score			0.037***						-0.007
			(0.005)						(0.004)
Quality: Citations				0.066***					0.016***
				(0.006)					(0.005)
Number of Collaborators					-0.006***				-0.012***
					(0.002)				(0.002)
Novelty						0.073***			0.010
						(0.006)			(0.006)
Topic Dispersion							-0.091***		-0.052***
							(0.003)		(0.003)
Seniority								0.052***	0.034***
								(0.004)	(0.004)
Seniority^2								-0.001***	-0.001***
								(0.000)	(0.000)
Observations	69020	69020	69020	69020	69020	69020	69020	69020	69020
Adj. R^2	0.046	0.029	0.006	0.017	0.000	0.014	0.029	0.018	0.107

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity measures are standardized within each exam. The sample is all candidates who applied for qualification in the 2012 wave of Abilitazione. Standard errors are clustered at the exam level.

change, the interaction between *Candidate's Quality of Publications* and *Committee Quality* is statistically not different from the results in the main text, and the point estimates only differ by 0.2 percentage points.

Column 3 reports results using an alternative measure of *Quantity*, scaling it by the average number of coauthors that the candidate has. This adjustment may be important if evaluators assign lower weight to publications that are produced by teams. We find that the R^2 is lower in this specification, implying that this variable is less in informative about the decision-making process of evaluators.

Next, we check whether our findings are robust to alternative measures of quality. First, we check whether the decision to use the share of high-impact articles is driving our results. Column 4 of Table A9 reports the results if we construct quality as the number of high-impact articles, rather than the share of high-impact articles. The key variables of interest

are statistically not different from the specification where we use the share of high-impact articles. Importantly, the marginal effect of publication quantity drops and marginal effect of publication quantity increases substantially. We believe this is because the quality measure also captures non-linearity in the quantity measure. Since the results are not driven by this decision, and econometrically we believe it is a better choice, in the main results we will be using the share of high-impact publications.

We also check alternative definitions of candidate quality. Columns 5 and 6 of Table A9 reports the results if we use the accumulated citations and average Article Influence Score of candidates as a measure of quality. We find that the same pattern holds as in the main text. The marginal effect of quality is positive and statistically significant. We observe that the interaction between *Candidate's Quality of Publications* and *Committee Quality* has a positive coefficient estimate but statistically it is not significant. In these specifications there is a negative and statistically significant estimate on the interaction between *Candidate's Number of Publications* and *Committee Quality*.

Table A9: Effect of committee composition on evaluation outcomes (alternative measures)

	Main-Text	OpenAlex	OpenAlex-Scaled	Count High-Impact	Citations	AIS
	(1)	(2)	(3)	(4)	(5)	(6)
Candidate Quantity Measure	0.117***	0.121***	0.112***	0.062***	0.104***	0.083***
	(0.005)	(0.004)	(0.004)	(0.006)	(0.005)	(0.006)
Candidate Quality Measure	0.098***	0.060***	0.064***	0.117***	0.021***	0.061***
	(0.006)	(0.004)	(0.004)	(0.008)	(0.004)	(0.008)
Committee Quality	-0.051*	-0.051*	-0.051*	-0.051*	-0.051*	-0.051*
	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)
Candidate Quantity Measure x Committee Quality	-0.025**	-0.008	-0.002	-0.040***	-0.032***	-0.038**
	(0.012)	(0.010)	(0.008)	(0.013)	(0.012)	(0.016)
Candidate Quality Measure x Committee Quality	0.032***	0.030***	0.028***	0.040**	0.000	0.023
	(0.012)	(0.010)	(0.010)	(0.016)	(0.010)	(0.019)
Observations	69020	69020	69020	69020	69020	69020
Adj. R^2	0.088	0.071	0.064	0.099	0.048	0.062
Exams	184	184	184	184	184	184

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. All candidate productivity measures are standardized within each exam, and evaluator quality is standardized within each eligibility pool before computing T_e. We instrument the final committee composition with the initial draw, before any resignations. All columns control for the expected committee quality. Standard errors are clustered at the exam level.

B.3 Alternative measures of evaluator quality

We construct a single measure of evaluator quality, defined as the number of high-impact articles. We consider an article high-impact if it is published in a journal that is in the top-quartile Web of Science journals for STEM+M fields or in a journal that is on the A-list of journals provided by the ministry in Social Science and Humanities.

Here, we check alternative measures of evaluator quality to verify that our results are robust to alternative definitions. Similarly to the section on candidate profiles, we define the following measures of committee quality:

1. *Quality/Impact (Journal-based)*: number of articles published either in the top quartile of the Web of Science journal rankings for STEM+M fields or in A-list journals designated by the ministry in Social Sciences and Humanities (measure used in the main-text)

2. *Quality/Impact (Article Influence Score)*: number of articles weighted by the article influence score of the journal

3. *Quantity in Web of Science*: number of articles published in journals tracked by the Web of Science by the evaluator

4. *Quantity*: number of articles published by the evaluator without adjusting for journal quality

5. *Quality/Impact (Citation-based)*: number of citations accumulated by the evaluator

Table A10 reports the results. Across all specifications, the coefficient estimate on *Committee Quality* is negative. With the exception of the citation-based measure all coefficients are statistically significant at the 10 percent level. The interaction between *Committee Quality* and *Candidate's Number of Publications* is negative and statistically significant across all specifications. The interaction between *Committee Quality* and *Candidate's Quality of Publications* is positive across all specifications, however statistically insignificant when we use *Articles in Web of Science* and *Citations* as a proxy for evaluator quality.

Table A10: Impact of committee composition on evaluation outcomes (alternative measures of committee quality)

	Main-Text	AIS	Articles WOS	Articles	Citations
	(1)	(2)	(3)	(4)	(5)
Candidate's Number of Publications	0.117***	0.116***	0.116***	0.117***	0.116***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)
Candidate's Quality of Publications	0.098***	0.099***	0.099***	0.098***	0.099***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)
Committee Quality	-0.051*	-0.051*	-0.049*	-0.047*	-0.018
	(0.030)	(0.026)	(0.028)	(0.027)	(0.032)
Candidate's Number of Publications x Committee Quality	-0.025**	-0.018*	-0.020*	-0.029**	-0.023**
	(0.012)	(0.011)	(0.011)	(0.012)	(0.010)
Candidate's Quality of Publications x Committee Quality	0.032***	0.024*	0.018	0.028**	0.010
	(0.012)	(0.013)	(0.013)	(0.014)	(0.013)
Observations	69020	69020	69020	69020	69020
Adj. R^2	0.088	0.088	0.088	0.088	0.086

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity measures are standardized within each exam, and evaluator quality is standardized within each eligibility pool before computing T_e. We instrument the final committee composition with the initial draw, before any resignations. All columns control for the expected committee quality. Standard errors are clustered at the exam level.

C Data

C.1 Merging of Datasets

We collect data from three different sources; the universe of full professors from CINECA, the process of national evaluations from ASN, and the research output of researchers from IRIS.

We use CINECA as the core of our database as it provides complete coverage of all full professors in Italian academia. We confirm that within the same university, field of study, and year there are no cases where we have two people with the same name.[18] Furthermore, there are very few cases where there is more than one professor with the same name in a given field of study.

We generate unique identifiers for all full professors in CINECA. Since we scrape multiple snapshots, we identify professors across multiple time periods. We do this in a three step procedure. For those professors that have a unique name across all time periods, we simply use their name to identify them across time periods. We assume that if a name appears across multiple time periods, it is the same person. For those professors whose names are not unique in all years, we use information on their area of study and affiliation. If a professor has a unique name within an area of study across all years, we assume that it is the same person. We assume that the area of study does not change for these individuals. This is not a restrictive assumption, as we are using an administrative database where each professor is assigned their field of study, that we believe is unlikely to be misreported or change over time. If there are multiple individuals with the same name, in the same field of study, we use their affiliations. For these (two) individuals we assume that their affiliations do not change over time, and create separate profiles for them at the (full name, area, university) level.

We start by merging ASN to CINECA. Since the ASN data is extracted from PDF files, we suspect that the affiliation is incorrectly assigned in certain cases. Hence, we adapt a three step merge algorithm. First, those authors that have a unique surname at the year-area level, we merge based on the following subset of variables (year, area, surname). Second,

18. There is one case in IRIS where we identify a single case of duplicated names. We manually confirm that this is a mistake in IRIS.

those authors that have a unique full name at the year-area level, we merge based on (year, area, surname, name). For the remaining few cases, where there are multiple researchers with the same name at the exam level, we also include university in the list of variables based on which we merge: (year, area, surname, name, university). This procedure yields a database of evaluators, where we identify all professors that participate in the national evaluations. We observe whether they were initially drawn to sit on the committee and whether they resigned.

Then we merge all professors in CINECA and our list of evaluators to IRIS. Here we are completely confident in the assignment of affiliation. We merge professors with names that do not appear more than once under the same affiliation in any period, we merge based on full name and affiliation. For those professors, whose names appear more than once under the same affiliation, we drop from our database. There was only one case where we observed such a duplicate, and we believe it was due to an error in the data.

D Randomization Checks

Table A11: Randomization Check on Candidate Characteristics

	(1) Publcations	(2) Share High-Impact	(3) Articles	(4) High-Impact	(5) AIS
Committee Quality (T_e)	-3.597	-0.014	-3.166	-2.398	-1.637
	(8.212)	(0.031)	(7.526)	(3.642)	(12.618)
Expected Committee Quality ($E[T_e]$)	34.585	-0.020	-7.555	-15.718	-78.525
	(68.809)	(0.386)	(61.710)	(36.285)	(130.655)
Observations	69020	67843	69020	69020	69020
R^2	0.001	0.000	0.001	0.002	0.001

Notes: The dependent variables are measures of candidate research productivity prior to the exam. *High impact articles* are candidates' articles in top-quartile Web of Science journals (STEM+M) or A-list journals (SSH). *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. Evaluator quality is standardized within each eligibility pool before computing T_e. We instrument the final committee composition with the initial draw, before any resignations. Standard errors are clustered at the exam level.

E Novelty

We adopt novelty measures developed in Arts, Melluso, and Veugelers (2025) to analyze whether better-published researchers value more novel research. We utilize the publicly released data file *papers_textual_metrics.csv* from https://zenodo.org/records/13869486. We merge the novelty measures to the publications of candidates in our database and construct multiple measures of novelty prior to the 2012 exams. We primarily use the ex-post measures that capture the introduction of new phrases, and word combinations to get as close as possible to the information set available to evaluators at the time of the exam.

We construct the following measures based on words at the candidate level:

1. *average new words:* the average of new words introduced in the candidate's papers
2. *average new word combinations:* the average of new combination of words introduced in the candidate's papers
3. *top new words:* the number of articles by the candidate that rank in the 95th percentile on new words on the sample of publications by all candidates
4. *top new word combinations:* the number of articles by the candidate that rank in the 95th percentile on new word combinations on the sample of publications by all candidates

We construct the following measures based on phrases at the candidate level:

1. *average new phrase:* the average of new phrases introduced in the candidate's papers
2. *average new phrase comb:* the average of new combination of phrases introduced in the candidate's papers
3. *top new phrase:* the number of articles by the candidate that rank in the 95th percentile on new phrases, on the sample of publications by all candidates
4. *top new phrase combinations:* the number of articles by the candidate that rank in the 95th percentile on new phrase combinations, on the sample of publications by all candidates

Table A12: Correlation between word-based novelty and qualification

	Overall				STEM				SSH			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average New Words	0.002				0.002				0.001			
	(0.002)				(0.003)				(0.005)			
Average New Word Combinations		0.005				0.008**				-0.004		
		(0.003)				(0.004)				(0.005)		
Top New Words			0.031***				0.043***				-0.009*	
			(0.004)				(0.004)				(0.006)	
Top New Word Combinations				0.035***				0.050***				-0.011
				(0.004)				(0.005)				(0.006)
Observations	69020	69020	69020	69020	41395	41395	41395	41395	27625	27625	27625	27625
Adj. R^2	0.129	0.129	0.131	0.132	0.130	0.130	0.136	0.138	0.126	0.126	0.126	0.126
Exams	184	184	184	184	109	109	109	109	75	75	75	75
Candidate Productivity Measures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity and novelty measures are standardized within each exam. Standard errors are clustered at the exam level.

Table A13: Correlation between phrase-based novelty and qualification

	Overall				STEM				SSH			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average New Phrases	0.010***				0.016***				-0.010**			
	(0.003)				(0.003)				(0.004)			
Average New Phrase Combinations		0.003				0.010**				-0.022***		
		(0.004)				(0.005)				(0.004)		
Top New Phrases			0.039***				0.054***				-0.009	
			(0.004)				(0.005)				(0.005)	
Top New Phrase Combinations				0.035***				0.051***				-0.014**
				(0.005)				(0.005)				(0.006)
Observations	69020	69020	69020	69020	41395	41395	41395	41395	27625	27625	27625	27625
Adj. R^2	0.129	0.129	0.132	0.132	0.131	0.131	0.139	0.138	0.126	0.127	0.126	0.126
Exams	184	184	184	184	109	109	109	109	75	75	75	75
Candidate Productivity Measures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity and novelty measures are standardized within each exam. Standard errors are clustered at the exam level.

Table A12 reports the correlation between qualification in the exams and the textual novelty measures. Curiously, we find that in STEM+M fields novelty is positively correlated with qualification, while in SSH there is a negative correlation. Table A13 reports a similar pattern for the phrase-based measures of novelty.

To make sure that this is not driven by systematic differences in abstract availability across disciplines, we redefine the measures to take into account the total number of words and phrases for each publication.

We construct the following measures based on words at the candidate level:

1. *average share new words:* the average of new words divided by total words
2. *average share new word combinations:* the average of new combination of words divided by total words
3. *top share new words:* the number of articles by the candidate that rank in the 95th percentile on new words divided by total words, on the sample of publications by all candidates
4. *top share new word combinations:* the number of articles by the candidate that rank in the 95th percentile on new word combinations divided by total words, on the sample of publications by all candidates

We construct the following measures based on phrases at the candidate level:

1. *average share new phrase:* the average of new phrases divided by total phrases
2. *average share new phrase comb:* the average of new combination of phrases divided by total phrases
3. *top share new phrase:* the number of articles by the candidate that rank in the 95th percentile on new phrases divided by total phrases, on the sample of publications by all candidates
4. *top share new phrase combinations:* the number of articles by the candidate that rank in the 95th percentile on new phrase combinations divided by total phrases, on the sample of publications by all candidates

Table A14: Correlation between adjusted word-based novelty and qualification

	Overall				STEM				SSH			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average Share New Words	-0.002				-0.002				-0.001			
	(0.002)				(0.003)				(0.004)			
Average Share New Word Combinations		0.005				0.011***				-0.011**		
		(0.003)				(0.004)				(0.005)		
Top Share New Words			0.032***				0.045***				-0.009*	
			(0.004)				(0.004)				(0.005)	
Top Share New Word Combinations				0.036***				0.051***				-0.007
				(0.004)				(0.004)				(0.007)
Observations	69020	69020	69020	69020	41395	41395	41395	41395	27625	27625	27625	27625
Adj. R^2	0.129	0.129	0.131	0.132	0.130	0.131	0.136	0.138	0.126	0.126	0.126	0.126
Exams	184	184	184	184	109	109	109	109	75	75	75	75
Candidate Productivity Measures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity and novelty measures are standardized within each exam. Standard errors are clustered at the exam level.

Table A15: Correlation between adjusted phrase-based novelty and qualification

	Overall				STEM				SSH			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average Share New Phrases	0.006**				0.007**				0.003			
	(0.002)				(0.003)				(0.004)			
Average Share New Phrase Combinations		0.009**				0.019***				-0.025***		
		(0.004)				(0.004)				(0.004)		
Top Share New Phrases			0.044***				0.058***				0.005	
			(0.004)				(0.004)				(0.007)	
Top Share New Phrase Combinations				0.040***				0.056***				-0.010
				(0.004)				(0.005)				(0.006)
Observations	69020	69020	69020	69020	41395	41395	41395	41395	27625	27625	27625	27625
Adj. R^2	0.129	0.129	0.134	0.133	0.130	0.131	0.140	0.140	0.126	0.127	0.126	0.126
Exams	184	184	184	184	109	109	109	109	75	75	75	75
Candidate Productivity Measures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity and novelty measures are standardized within each exam. Standard errors are clustered at the exam level.

Tables A14 and A15 reports similar patterns as before. The correlation between qualification in the exams and the textual novelty measures is positive overall and in STEM+M, in SSH we observe is a negative correlation.

We also explore whether the measure *semantic_distance* yields results that are comparable

across broad discipline categories. We take the average semantic distance of candidates' publication and correlate it with whether they qualify. Table A16 shows that this measure does not predict qualification in our data.

Table A16: Correlation between semantic-distance-based novelty and qualification

	Overall	STEM	SSH
	(1)	(2)	(3)
Semantic Distance	0.002	0.000	0.006
	(0.003)	(0.004)	(0.005)
Observations	69020	41395	27625
Adj. R^2	0.129	0.130	0.126
Exams	184	109	75
Candidate Productivity Measures	✓	✓	✓
Candidate Connections	✓	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity and novelty measures are standardized within each exam. Standard errors are clustered at the exam level.