Top Researchers on Scientific Committees: Decision Quality, Peer Effects, and Opportunity Costs

Milan Makany Natalia Zinovyeva*

July 15, 2026

Click here for the latest version

Abstract

Science disproportionately relies on top researchers to evaluate the work of others, potentially diverting their scarce time from research, mentoring, and other service. We study this trade-off in Italy's national academic qualification system, where evaluators are randomly assigned to committees. Committees with better-published evaluators select candidates with stronger subsequent citation and career outcomes; they also place greater weight on publication impact rather than quantity. These effects partly reflect spillovers within committees: evaluators assigned to serve alongside more accomplished colleagues exert more effort and adjust their criteria. Assignment to a committee, however, reduces evaluators' subsequent publication output and PhD supervision, especially among top researchers. Research losses are especially large in the Social Sciences and Humanities, and among those working in small teams. More productive researchers are also disproportionately less likely to volunteer for service again.

JEL Codes: I23, D71, D83, M51, J45

Keywords: expert evaluation, committees, academic promotions, decision quality, peer effects, opportunity costs

*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, Igo Dobbe, Patricia Himml, Sacha Kapoor, Attila Lindner, Jacques Mairesse, Otto Swank, Reinhilde Veugelers, Bauke Visser, 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 Birmingham; University of Edinburgh; University of Warwick; Erasmus School of Economics; University of Ottawa; Aalto University; SOFI; and CEBI at the University of Copenhagen. Natalia Zinovyeva gratefully acknowledges support from the CAGE Research Centre, funded by the ESRC.

1 Introduction

Scientific evaluation by peers is at the core of scientific progress. By allocating publication, funding, promotion, and recognition, peer judgment helps shape the direction of science and the incentives to produce high-quality research. Scientific evaluation, however, not only helps allocate scarce scientific resources; it also uses scarce scientific capacity. Expert evaluation requires substantial amounts of time from researchers whose time is scarce and whose contributions may be especially valuable in research, mentoring, and other service. The journal peer-review system alone is estimated to absorb more than 100 million researcher-hours annually, with much of the burden concentrated among a relatively small group of highly qualified researchers.[1] This burden appears increasingly difficult to sustain, as scientific organizations report growing difficulty recruiting qualified evaluators.[2]

Despite its importance, we know surprisingly little about either side of this trade-off. It remains unclear whether more accomplished evaluators make better decisions and whether, in collective evaluations, their presence improves outcomes by altering the behavior of other committee members. Even less is known about the opportunity cost of evaluation service. These costs may be particularly large when service diverts time from highly productive researchers, and they may extend beyond evaluators' own output: joint projects may slow, the production of new knowledge may be delayed, and students may receive less supervision. Participation may also generate a dynamic cost if serving today reduces researchers' willingness to undertake evaluation duties in the future.

In this paper, we provide causal evidence on the benefits and costs of assigning better-published researchers to scientific evaluation committees. We ask whether evaluators with stronger publication records improve committee decisions; whether their presence shapes the behavior of other committee members; and what research, training, and future service are displaced when they are assigned to serve. We study these questions in Italy's national academic qualification system, the *Abilitazione Scientifica Nazionale*, where candidates for Associate and Full Professor positions are evaluated by discipline-level committees. After candidates apply and eligible professors volunteer for service,

1. Aczel et al. (2021) estimate that peer review absorbed more than 100 million researcher-hours in 2020 across major countries, worth over $2.5 billion when valued at average hourly wages. The workload is highly concentrated: 10% of reviewers perform 50% of reviews (Publons, 2018; Vesper, 2018), consistent with accounts of heavily solicited researchers cutting back on reviewing (Dance, 2023, for example, on *Nature Careers*).

2. *Nature*'s coverage of the Publons Global State of Peer Review, based on a survey of more than 11,000 researchers, reports 'reviewer fatigue," difficulty finding willing reviewers, and rising invitations per completed review (Publons, 2018; Vesper, 2018). A UKRI-commissioned review similarly highlights reviewer burden, weak incentives, and the need to sustain reviewer supply for grant evaluations(Kolarz et al., 2023). In April 2026, the ERC tightened resubmission rules for its 2027 competitions, explicitly citing rising application volumes and pressure on reviewers, before partially revising the rules less than two weeks later following concerns from the research community (ERC, 2026). Surveys also identify time burden as a major obstacle to recruiting and retaining editorial board members (Funk et al., 2024).

committee members are randomly drawn from the pool of potential evaluators subject to known institutional constraints.[3] This random assignment allows us to address all three questions causally. First, we study selection quality by comparing outcomes of candidates evaluated by committees whose members have, by luck of the draw, stronger or weaker publication records. Second, we study peer dynamics – using individual votes and written reports – by exploiting the fact that, conditional on being drawn, the quality of one's committee peers is itself randomly assigned. Third, we estimate the opportunity costs of service by comparing professors randomly drawn to committees with otherwise similar professors who were not.

Our database covers around 10,000 eligible professors who volunteered, of whom about 3,200 served on more than 680 committees between 2012 and 2021. For the first national round, we observe all applications ($\approx 60,000$), across the system's 184 disciplinary areas, as well as roughly 300,000 individual evaluation reports covering each evaluator-candidate pair. We link these administrative records to bibliographic data and PhD dissertation records to follow candidates' careers, and evaluators' research and PhD student supervision.

We first show that better-published committees are stricter and use different evaluation criteria. Replacing an average evaluator with one whose pre-committee publication record is one standard deviation stronger lowers candidates' probability of qualification by 1.5 percentage points, or about 4 percent relative to the 37 percent baseline. Better-published committees also place different weights on candidates' observable characteristics: their selection is associated less with high publication quantity and more with high publication impact.

These stricter standards and different criteria are accompanied by better selection. Using decade-long follow-up data, we compare future outcomes of candidates assessed by committees with different composition. To distinguish better selection from greater strictness, we examine qualified and non-qualified candidates separately: if better-published committees were simply raising the qualification threshold, their rejected candidates should also have stronger future outcomes. Instead, we find that they select such that the future citations of qualified candidates are greater, and that of rejected candidates are smaller, for both works published prior to the evaluation and

3. As we explain in more detail in Section 2, since 2012, candidates for Associate and Full Professor positions at Italian public universities must obtain a national qualification before they can be considered for promotion. Candidates apply for evaluation by discipline-level committees in response to public calls. In each evaluation cycle (2012, 2016, 2018, 2021), committees are drawn after the first wave of applications are submitted and serve for two years, evaluating candidates from the first wave and subsequent waves. Each committee consists of five Full Professors from the corresponding disciplinary area. For example, within economics there are separate committees for economic history (*storia economica*), econometrics (*econometria*), and public economics (*scienza delle finanze*). Candidates are evaluated based on their submitted curriculum vitae and publications.

works produced after. Candidates qualified to Associate Professor by better-published committees are also more likely to become Full Professors within ten years. Specifically, if an average committee member is instead replaced by another who is one standard deviation better published, it leads roughly a one percentage point increase in promotion probability, or 6 percent relative to the 17 percent baseline. with no corresponding difference among non-qualified candidates. Similarly, their selected candidates are also more likely to enter the future pool of potential evaluators. These differences persist after controlling for candidate characteristics, including past research productivity and affiliation. The results indicate that evaluator expertise not only shifts the location of the qualification threshold, but also improves ranking.

We then ask whether better-published researchers affect committee decision-making only through their own votes or also through influence on their peers. We find that better-published evaluators are themselves less likely to cast positive votes and place greater weight on publication impact. Evaluators randomly assigned to more accomplished peers adopt similar standards and devote more effort to their reports. A one standard deviation increase in the publication record of a peer raises the length and lexical complexity of an evaluator's written reports by about 7 percent of a standard deviation. We find no evidence that better-published evaluators themselves exert less effort while serving. We also find suggestive evidence that the effects of evaluator expertise are also stronger at the extensive margin: point estimates suggest larger gains from adding the first better-published evaluator to a committee than from further increases in committee expertise.

The benefits of expert involvement come with measurable costs and externalities. Serving on an ASN committee requires evaluating hundreds of candidates over a two-year term in addition to regular academic responsibilities. Random assignment to committee service reduces publication output during and after service by about 5 percent of a standard deviation per year, cumulating to roughly 20 percent of a typical year's publications.[4] Committee service also reduces PhD supervision, with effects emerging several years after the draw, consistent with evaluators taking on fewer new students while serving. The burden of service is therefore not fully absorbed through short-run adjustments in effort.

These costs depend on the organization of scientific production. Publication declines are larger in Social Sciences and Humanities, particularly among researchers who tend to work in smaller teams. Researchers who work alone may be less able to rely on collaborators when committee

4. In economics, this corresponds to about 0.5 fewer publications over the three years following the start of the evaluation process.

service disrupts their research time. Because production in larger teams and labs is more prevalent in STEM fields, the heterogeneity by collaboration patterns likely reflects a core feature of the scientific production function. By contrast, disciplinary differences in opportunity costs do not appear to be driven primarily by differences in the report-writing burden associated with evaluation.

Research-output losses are also larger among more productive researchers. Part of this difference reflects their higher baseline output: a similar proportional disruption generates a larger absolute loss when applied to a more productive researcher. But such proportional losses were not inevitable. More accomplished researchers might also be more efficient evaluators, completing the same task with less time or attention and therefore sacrificing a smaller share of their research output. Instead, we find no evidence that they economize on evaluative effort – their reports are no shorter or less detailed – and their research output falls proportionally by at least as much as that of other evaluators. Since these researchers account for a disproportionate share of scientific production, their assignment to committee service consequently produces especially large losses in knowledge output.

We also find evidence that service affects the future supply of expert evaluators. For researchers whose publication records are one standard deviation above the pool average, the negative effect of random committee assignment on subsequent volunteering is nearly twice as large as for the average researcher. Aggregate participation patterns over the decade following the introduction of the two-tier promotion system point in the same direction: top researchers become progressively less likely to serve. These findings suggest a potential dynamic cost of repeatedly drawing on scarce expertise, although we interpret them cautiously because subsequent participation may also be shaped by institutional features of the selection process.

Taken together, the results show that scarcity of expert evaluators is not only an exogenous constraint faced by evaluation systems; it can also be a consequence of how those systems use expert time. Better-published evaluators improve selection, shape the behavior of their peers, and select candidates who are more likely to strengthen the future senior pool of evaluators. At the same time, service reduces research output, affects students and collaborators, and lowers future willingness to serve, especially among the most productive researchers. Voluntary participation does not ensure that these costs are fully taken into account, because evaluators may not internalize the effects of their service on others. Then, the design problem is to allocate expert attention where its marginal value is high while avoiding service burdens that erode the stock of expertise on which evaluation systems depend.

This paper makes several contributions to the literature. First, we contribute to work on the organization of science by studying peer evaluation not only as a mechanism for allocating scientific rewards, but also as an activity that uses scarce scientific labor. A long tradition emphasizes that peer judgment is central to the allocation of credit, resources, and careers in science (Merton, 1957; Stephan, 2010). Recent debates often ask how far such judgment can or should be supplemented by bibliometric indicators, which may reduce information frictions (Hager et al., 2024) but also risk replacing expert assessment with imperfect mechanical rules (DORA, 2012; Hicks et al., 2015; Bertocchi et al., 2015; Stephan et al., 2017; Heckman and Moktan, 2020). Our results show why this trade-off is incomplete. Even in a setting with rich bibliometric information, better-published evaluators select candidates with stronger future outcomes and who are more likely to become highly qualified evaluators in the future, indicating that expert judgment does add information beyond observable indicators. But this judgment is costly to produce: service reduces research output, affects training, and lowers future participation by the very researchers whose evaluations are most informative. Thus, the capacity of science to evaluate itself depends not only on how candidates are assessed, but also on whether evaluation systems preserve the scarce expert time on which peer judgment relies.

Second, we extend research on how evaluator characteristics shape committee decisions. Prior work has examined the roles of gender (De Paola and Scoppa, 2015; Bagues et al., 2017; Card et al., 2020; Funk et al., 2024), connections and favoritism (Brogaard et al., 2014; Zinovyeva and Bagues, 2015; Bagues et al., 2019a), and academic proximity between candidates and evaluators (Li, 2017; Krieger et al., 2023). Much less is known about how evaluators' own research excellence affects evaluation outcomes. Existing evidence suggests that less-qualified evaluators may rely on noisier signals of quality (Bagues et al., 2019b), but systematic causal evidence on the role of evaluator excellence is lacking. We show that better-published evaluators systematically apply stricter standards, shift emphasis from publication quantity to impact, and select candidates with stronger future outcomes. Our results complement prior evidence that subject-matter experts without close ties improve selection outcomes (Zinovyeva and Bagues, 2015; Li, 2017), by highlighting a distinct dimension of expertise: evaluators' research excellence.

Third, we contribute to the literature on committee decision-making and peer effects in professional settings. Theoretical work on information-sharing committees emphasizes that the presence of informed members can change others' incentives (Ottaviani and Sørensen, 2001), and that career and reputation concerns may alter behavior in collective decisions (Prat, 2005; Levy, 2007; Visser

and Swank, 2007; Swank and Visser, 2023). We provide causal evidence from a high-stakes professional setting in which committee composition changes both evaluation standards and effort. Peers serving with top researchers become stricter, place more weight on research impact, and write more demanding reports. This pattern is consistent with within-committee reputation concerns, even in a setting where individual evaluations and votes are public. These findings also complement evidence from personnel economics showing that coworkers' productivity affects individual effort (e.g., Chan, 2021).

Fourth, we advance understanding of the opportunity costs of academic service. Peer review and committee work are recognized to require substantial, and often uncompensated, researcher time. Yet existing evidence has relied primarily on self-reported survey data. Exploiting random assignment to evaluation duties, we provide causal evidence that committee service generates large and persistent reductions in researchers' subsequent publication output. We further show that these costs extend beyond researchers' own publications: service also reduces PhD supervision. This finding is consistent with evidence that faculty availability is an important input into graduate training and students' subsequent careers (Waldinger, 2010). We also show that the output costs of service depend on the mode of scientific production. Publication losses are larger in Social Sciences and Humanities, particularly among researchers who typically work in smaller teams. These patterns suggest that these costs depend on how scientific production is organized and on the ability of collaboration networks to partly absorb disruptions to an individual researcher's time. Losses are also larger in absolute terms among more productive researchers. Although this partly follows from their higher baseline output, it makes the allocation of service consequential for aggregate knowledge production: diverting the time of researchers who account for a disproportionate share of scientific output creates larger external costs for science and its users. These findings contribute to research on multitasking and incentives (Holmstrom and Milgrom, 1991) and to recent work on the allocation of effort and incentives in academia (Azoulay et al., 2011; Checchi et al., 2021; Nieddu and Pandolfi, 2022; Myers and Tham, 2023; Hoffman and Stanton, 2024).

The remainder of the paper is organized as follows. Section 2 describes the institutional setting of Italian academic promotions. Section 3 outlines our data sources. Section 4 presents results on selection outcomes and costs, and discusses the mechanisms through which better-published researchers affect decision-making. Section 5 concludes with implications for policy and theory.

2 Italian National Academic Qualifications

Italy's *Abilitazione Scientifica Nazionale* (ASN) was a national qualification system introduced in 2012 as part of the Gelmini Reform (Law 240/2010). The ASN was a mandatory first-stage filter for promotions at public universities: candidates for Associate or Full Professor had to first obtain a national qualification from a discipline-level committee before they could apply for promotion at the university level. The system was designed to harmonize standards across universities and limit local favoritism.

Three institutional features are central to our design. First, in the first ASN wave, candidates applied before learning the identity of the committee that would evaluate them. Second, committee members were selected by public random draw, subject to a small set of institutional constraints. Conditional on these constraints, the draw generated random variation in committee composition, in the peer quality faced by each evaluator, and in whether eligible volunteers served. Third, committees retained discretion in defining and applying evaluation criteria, allowing us to study whether better-published committees use different criteria and select candidates with stronger subsequent research and career outcomes.

2.1 Public call and applications

The ASN was organized through field-level public calls. Candidates applied for qualification as Associate or Full Professor in a specific discipline and could apply in multiple fields. In the first wave of each cycle, applications were submitted before candidates observed committee composition. Candidates could therefore choose whether and where to apply, but they could not condition their application decision on the realized composition of the committee.

The process was highly transparent. Evaluators' and candidates' curricula vitae, individual votes, and evaluation reports were published online. The national evaluation agency also collected and published information on candidates' research output, comparing each candidate's record with discipline-specific benchmarks for the rank sought.

2.2 Evaluator Selection

Evaluation committees were formed at the field-level and consisted of five members.[5] Committees were formed in three steps. First, the ministry defined a pool of professors who satisfy field-

5. Official documents regulating the process are available at https://abilitazione.mur.gov.it/public/index.php?lang=eng (retrieved September 2025).

specific research productivity thresholds. Second, eligible professors decided whether to volunteer for committee service. Third, committee members were selected from the resulting volunteer pool through a public random draw, subject to institutional constraints.

Eligibility. Eligibility criteria reflected disciplinary publication norms. In STEM fields (including medicine and psychology), professors were eligible if their research output exceeded field-specific thresholds based on journal publications, citations, and the H-index. In the Social Sciences and Humanities (SSH), eligibility was based instead on journal publications, articles in high-impact journals, and the number of books and book chapters. These thresholds determined the set of professors who could enter the evaluator pool within each field. The thresholds themselves slightly changed across cycles.

Volunteering. Eligibility did not automatically place professors in the pool of potential evaluators. The ministry first issued a public call for volunteers. Volunteers were asked to report their university, discipline, and provide their curriculum vitae. The ministry then verified their eligibility and determined whether they could serve.

Draw procedure. The draw of committee members was implemented centrally by the Ministry through a public computerized procedure.[6] Within each field, eligible volunteers were ordered alphabetically and assigned an identifier. The Ministry then generated a random sequence without replacement and applied this sequence to the field-specific lists to determine the order in which evaluators were selected.

Committees were filled sequentially according to this randomized order, subject to two institutional constraints. First, no more than one evaluator from the same university could serve on the same committee. If an assignment would violate this constraint, the procedure moved to the next eligible name in the randomized list. Second, in fields that contained multiple large sub-fields, the algorithm imposed sub-field representation rules before filling the remaining seats.

If a selected evaluator resigned or could not serve, a replacement was drawn according to the same randomization procedure.

Conditional on the field-specific pool of potential evaluators and these constraints, the initial composition of committees was random. This random variation allows us to identify the causal

6. Official documentation is available for the 2012 cycle at https://abilitazione.mur.gov.it/public/documenti/Modalita_sorteggio.pdf and for later cycles at https://abilitazione.mur.gov.it/public/documenti/2016/Allegato_1_Procedura_sorteggio_commissioni_ASN_2016.p (accessed June 27, 2025).

effect of committee composition, peer effects, and the costs of service.

In the 2012 ASN cycle, four members were drawn from the pool of eligible Full Professors affiliated with Italian universities, while the fifth member was drawn from a separate list of researchers based in OECD countries outside Italy. The OECD member was selected through an analogous randomization procedure. This foreign-member requirement was removed in later cycles, in which all five members were drawn from lists of Full Professors serving at Italian universities. Through out the paper we will be reporting results, excluding the foreign committee members, and focusing on the Italian evaluators.[7]

Mandate and service conditions. Committee members served for approximately two years and evaluated all applications submitted during their term.

After completing a term, evaluators were ineligible to serve again for at least three years, which affects our analysis of future volunteering outcomes in Section 4.2.

For Italy-based evaluators, service was essentially uncompensated: official rules provided no remuneration, although evaluators could request partial teaching-load reductions. In practice, committee service therefore represented a largely pro bono academic duty.

2.3 Evaluation Process

After the appointment of evaluators, committees had an initial meeting in which members defined the organizational rules and evaluation criteria for the procedure, before accessing candidates' applications. In the 2012 cycle, committees had substantial discretion in setting evaluation standards.[8] The committee members then individually evaluated each application on the basis of their criteria and produced individual judgments by each evaluator. These evaluations were discussed and finalized in subsequent committee meetings, which also recorded the members' votes and committee's final decisions. Final decisions required a qualified majority of at least four out of five votes in favor.

Evaluations were based exclusively on the material submitted by candidates: their curricula

7. Contemporary commentary suggested that the requirement of a foreign member created practical and legal difficulties, including higher dropout rates, uncertainty about equivalence with Italian ranks, and language barriers in some fields (e.g., see ROARS, 2014). According to our data, foreign evaluators wrote, on average, 20% shorter reports, consistent with their lack of engagement. All results are robust to including foreign evaluators, though the estimates are noisier.

8. The Ministry recommended that they take into account three bibliometric indicators similar to those used to determine evaluator eligibility. In later cycles, minimum bibliometric conditions were formally introduced, while committees retained discretion over additional criteria and the overall assessment of candidates.

vitae and up to ten selected publications.

The workload was substantial: based on informal discussions with evaluators across fields and cycles, service required roughly 200–300 hours over the two-year term, covering not only the review of individual files but also coordination meetings, the elaboration of evaluation criteria, writing individual reports for each candidate, and collective committee deliberations. Between 2012 and 2018, each committee evaluated on average around 270 candidates (Italian National University Council, 2023).

Following the evaluation, individual reports, committee summaries, and final decisions are published online. The individual reports were a central component of the evaluation process and also represent an important feature of our analysis. Evaluation reports provide a direct proxy of evaluator effort and allow us to study whether better-published committee members affect not only final decisions, but also the criteria and intensity of evaluation within committees.

3 Data

We compile a dataset linking candidates, evaluators, committee assignment, and outcomes in the ASN. We merge these records with administrative data on Italian professors, several bibliographic indexes, and a repository of doctoral dissertations. This section describes the underlying data sources and the main variables used in the analysis.

3.1 Data sources

Administrative data on Italian professors. We collect administrative records on all professors employed at Italian public universities between 2010 and 2023 from the Ministry of Universities and Research.[9] The resulting panel covers approximately 90,000 professors across 95 public universities and reports affiliation, department, research area, academic rank, and gender.

National qualification system (ASN). We collect administrative data on all researchers who volunteered as evaluators in ASN.[10] These records include (i) the pool of eligible evaluators, (ii) evaluators initially drawn by lottery, and (iii) the evaluators who ultimately served after resignations and substitutions. The dataset spans four ASN evaluation cycles conducted between 2012 and 2021.

9. The data are available at https://cercauniversita.mur.gov.it/php5/docenti/cerca.php, last accessed 15 November 2024.

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

For the first wave of the 2012 cycle, we further use the database assembled contemporaneously by Bagues et al. (2017) from the Ministry website.[11] These materials contain the names and curricula vitae of all pre-registered candidates, as well as the individual and committee evaluation reports posted as outcomes of the process. The reports allow us to observe each evaluator's individual vote on each candidate and to construct text-based proxies for evaluator effort, including report length and linguistic complexity.

Bibliographic databases. We combine three sources to measure candidates' and evaluators' research productivity before and after the qualification exams.

First, we extract candidates' publication records from the curricula vitae submitted with ASN applications. These records mirror the information available to evaluators at the time of assessment and allow us to reconstruct candidates' observable research output.

Second, for both candidates and evaluators, we compile publication and citation histories from IRIS and OpenAlex. IRIS, the Institutional Research Information System managed by ANVUR, contains self-reported publication records for researchers at Italian public universities and research centers. The Ministry uses these records in national research assessments that affect grant allocation and university funding. IRIS includes approximately eight million publications from 2000 to 2024.[12] We match 80% of all professors and 88% of potential evaluators to records in IRIS. To mitigate concerns arising from selective self-reporting, we supplement IRIS with OpenAlex, an open-access bibliometric index covering over 200 million works and approximately 90 million researchers worldwide (Priem et al., 2022). We match 89% of professors in the administrative database and 90% of potential evaluators to OpenAlex records.

We merge IRIS and OpenAlex by de-duplicating on title, publication year, and journal. When a publication appears in both sources, we retain citation counts from OpenAlex. We link publication data to administrative records by matching names, affiliations, and research fields; Appendix B describes the full merge procedure. Together, IRIS and OpenAlex cover 96% of professors in our administrative dataset.

Doctoral graduates database. We measure doctoral supervision using records on PhD graduates from the National Central Library of Florence (BNCF). In Italy, doctoral theses are subject

11. Candidates' curricula vitae, evaluation reports, and the names of unsuccessful candidates are available online for only six months after the decision.

12. We collected information from universities' IRIS portals in December 2024. These portals share a similar interface (e.g., Sapienza University of Rome: https://iris.uniroma1.it/, last accessed September 2025).

to legal deposit at the National Central Libraries of Florence and Rome.[13] In practice, the digital visibility of theses in the libraries' online catalogs depends on the timing and modality through which universities implemented systematic electronic deposit and harvesting procedures. As a result, coverage in the BNCF catalog is uneven in earlier years and becomes systematic only once institutional workflows stabilize.

We therefore interpret the Florence records as capturing the universe of doctoral theses that are digitally deposited and catalogued at BNCF, rather than the full legal universe of Italian doctoral theses. In the empirical analysis, we focus on the post-2012 period, when coverage becomes substantially more complete, and restrict attention to institutions with stable reporting. Specifically, we retain institutions with at least five recorded doctoral theses per year per 100 tenured faculty members in 2012. This threshold is empirically motivated: institutions above it display substantially more regular thesis deposits in subsequent years, while many institutions below it have sporadic or near-zero records. The resulting sample has much denser reporting. Among institutions above the cutoff, the median annual reporting rate over 2012–2023 is 17.6 theses per 100 tenured faculty members, and the 25th percentile is 9.8. By contrast, among institutions below the cutoff, the median is only 1.3, and the 25th percentile is 0.5, inconsistent with known PhD activity and likely reflecting incomplete catalog coverage. Applying this criterion excludes approximately 43% of institutions.

3.2 Descriptive statistics

Evaluators. Participation as an evaluator required satisfying eligibility thresholds and volunteering for service. Figure 1 describes these two margins.[14] The eligibility thresholds screened professors substantially with about half of Full Professors being eligible in most broad field groups, outside the 83% SSH eligibility rate in 2012 cycle. Conditional on eligibility, volunteering was also substantial. In 2012, about half of eligible professors volunteered for service. This willingness declined over time: by 2021, the volunteering rate had fallen to roughly 35%.

Across the four examination cycles, there were 685 evaluation committees in 184 fields, with around 10,000 Full Professors who both satisfied the eligibility requirements and volunteered for service. The average pool of potential evaluators contained 26 professors. In the 2012 round, committees included one foreign-based evaluator, so only four of the five members were drawn

13. See the BNCF documentation: https://bncf.cultura.gov.it/servizi/deposito-delle-tesi-di-dottorato/.

14. Because actual eligibility as determined by the Ministry is not directly observed, we reconstruct eligibility in our data using the official Ministry criteria.

from Italy-based volunteers.[15] In our analysis, we focus on Italian evaluators and exclude foreign members.

The evaluator pool was positively selected relative to the population of Full Professors. Eligible professors had about 70% of a standard deviation more total research output than ineligible professors (Appendix Table A1, Panel A). Selection further operated through volunteering: among eligible professors, volunteers had stronger publication records than non-volunteers (Appendix Table A1, Panel B).

This selection matters for interpreting both the benefits and costs of committee service. The effects of evaluator quality are identified within a pool of professors who are already positively selected relative to the population in their field. If better-published evaluators improve decisions even within this selected pool, the impact of expert participation may be even larger in settings where committees are drawn from a less selected group. At the same time, the costs of service are borne by this same high-output group, making our estimates of productivity losses especially relevant for informing systems that rely on scarce expert time.

Around 9% of initially drawn evaluators resigned before serving and were replaced through the random replacement procedure. Resignations are only weakly related to observable publication records (Appendix Table A1, Panel C), but they can nevertheless affect realized committee composition. In the empirical analysis, we therefore instrument realized serving-committee quality with the exogenous quality shock from the initial draw, addressing potentially endogenous resignations.

The publication-based measures used to characterize evaluator and committee quality are constructed from the ten years preceding each exam. The average researcher in the pool of potential evaluators had 60 publications in this pre-exam window, about 90% of which were journal articles; the remainder consisted of books, chapters, conference proceedings, and other outputs (Appendix Table A2). Roughly 40% of journal articles appeared in high-impact journals. We define high-impact journals using field-specific classifications: Q1 journals by Article Influence Score in STEMM fields, and the Italian evaluation agency's list of A-journals in economics, business, and other SSH fields.[16] This hybrid definition improves comparability across fields, since continuous impact metrics are not consistently available in SSH.

Our baseline measure of evaluator quality is the number of high-impact journal articles published

15. In 2012, across 20 disciplinary areas, too few foreign evaluators volunteered. In these areas, all five committee members were therefore drawn from Italian evaluators.

16. AIS is similar to a 5-year impact factor but weights citations by the influence of the citing journal and by the inverse number of references, and excludes self-citations.

in the ten years preceding the exam. This measure combines output volume and journal impact, aligns with evaluation practices in Italy, and has a direct policy interpretation. To account for cross-field differences in publication norms, we normalize this measure within evaluator-pool cells (discipline × year). We also define a binary measure, *top researchers*, equal to one for professors in the top 25% of all Italian Full Professors within the same discipline by this high-impact publication measure. In the pool of potential evaluators, *top researchers* produced 84 publications on average in the preceding ten years, compared with 45 publications among other researchers (Appendix Table A2).

We measure *committee quality* as the average normalized quality of the evaluators assigned to a committee. This continuous measure captures the aggregate publication strength of the committee and is the main source of variation used to study how evaluator quality affects committee decisions.[17]

Finally, we track evaluators' post-assignment productivity to measure the opportunity cost of committee service. We observe research output in the year of committee assignment and in subsequent years. In the four-year window starting with the assignment year, the average researcher in the pool of potential evaluators produced 28 publications, roughly one third of which were high-impact articles, and supervised 2 PhD graduates (Appendix Table A3). These outcomes allow us to compare the subsequent research output of researchers assigned to committee service with those potential evaluators in the pool who could have been drawn but were not.

Candidates. The first wave of the 2012 ASN cycle was exceptionally large. There were 69,020 pre-registered applications, corresponding to about 375 applications per field. More than 46,000 researchers applied, equal to roughly 61% of Assistant Professors and 60% of Associate Professors employed in Italian universities at the end of 2012.[18] About one third of applicants submitted multiple applications, either across related fields or across ranks within the same field.

Candidates could withdraw after committees were drawn and evaluation criteria were announced. The database from Bagues et al. (2017) includes the curricula vitae of all pre-registered candidates, posted online before the final list of evaluators was confirmed. Roughly 14% of pre-registered applications were withdrawn, leaving 59,151 applications to be evaluated (Italian National University Council, 2023). Withdrawals were not systematically related to assigned commit-

17. Appendix C examines robustness to alternative measures used to characterize candidate and evaluator quality and to measure candidates' subsequent outcomes. Appendix C.5 focuses specifically on alternative measures of evaluator and committee quality.

18. This figure is based on our calculations using Ministry records on all assistant (*Ricercatori*) and Associate Professors (*Associati*) as of December 31, 2012.

tee quality (Appendix Table A4). Overall, 37% of pre-registered applications (43% of evaluated applications) were successful. Two thirds of applicants were already employed at an Italian university, typically in a permanent position in the field in which they sought qualification.[19]

To study how evaluator expertise shapes evaluation criteria, we summarize candidates' pre-evaluation research profiles along two dimensions: *quantity* and *impact*. We measure *quantity* as the total number of academic outputs listed on candidates' curricula vitae, including journal articles, books, book chapters, conference proceedings, and patents. We measure *impact* as the share of journal articles published in high-impact journals. In science, technology, engineering, mathematics, and medicine (STEMM fields) high-impact journals are journals in the top quartile of Web of Science by Article Influence Score. In economics, business, and other social science and humanities (SSH fields) continuous journal metrics are less consistently available, hence we use the Italian evaluation agency's list of A-journals. Because there is no comparable impact metric for non-journal outputs, the impact measure is restricted to journal articles. Defining impact as a share also avoids publication impact from capturing non-linear effects of publication quantity.[20]

Publication practices differ sharply across fields (Appendix Table A5). In STEMM fields, the mean share of top-quartile articles ranges from 30% to 50% across fields. In economics and business, the mean share is about 15% using Q1 journals and 21% using A-journals; in the remaining SSH fields, the mean share of A-journal articles is about 29%. Given this heterogeneity across fields, we normalize both quantity and impact within evaluation fields.

We use candidates' curricula vitae to measure pre-evaluation research profiles because these were the records available to committees at the time of evaluation. We then validate these CV-based measures against the combined *OpenAlex+IRIS* publication database. In *OpenAlex+IRIS*, the average candidate has 29 matched publications, of which 28 are journal articles; 43% of these journal articles appear in high-impact journals. The correlation between the number of journal articles in candidates' curricula vitae and in *OpenAlex+IRIS* is 78%, suggesting that the combined external database captures journal-based output reliably. Correlations are lower for books and chapters, which are less consistently indexed in external publication databases.[21]

19. Among applicants for Associate Professor, 42% held tenured Assistant Professorships (*Ricercatore a tempo indeterminato*).

20. Appendix C.2 shows that these measures have the strongest predictive power for evaluation outcomes compared to citations, average AIS, number of collaborators, seniority, topic diversity, and novelty. Appendix C.4 shows robustness to alternative definitions of quantity and impact.

21. Appendix Tables C1 and C2 report descriptive statistics and correlations. Appendix Table C6 shows robustness to the choice of data source for candidates' research output.

When analyzing the effect of committee composition, we also control for candidate–evaluator proximity. About 15% of candidates had either a co-author or a colleague from their institution on their committee.[22] For candidates employed at Italian universities, we observe both the field (*settore concorsuale*) and subfield (*settore scientifico-disciplinare*) of appointment. Information on subfields allows us to measure subfield proximity between candidates and evaluators.

Finally, we track candidates' post-evaluation outcomes using publication data from *OpenAlex+IRIS* and career records from CINECA. These outcomes are used to assess whether committees with stronger publication records select candidates whose later performance is stronger. We measure future scientific output after the committee decision, including future citations received, distinguishing between citations to work published before the evaluation from citations to work published after the evaluation. On average, successful candidates produced 62 publications in the 10 years following the year of evaluation (Appendix Table A6). Their pre-evaluation publications received about 1,000 citations after the qualification decision, while publications produced after the evaluation received about 1,500 citations. We measure career progression using subsequent appointments in Italian universities. Among candidates qualified for Associate Professor, 68% were promoted to Associate Professor and 17% to Full Professor within ten years. Among candidates qualified for Full Professor, 57% attained a Full Professorship within ten years.

Individual evaluation reports. Each committee member wrote an individual report for every application. Reports summarize the candidate's research profile, provide a qualitative assessment, and conclude with a recommendation on qualification. The dataset contains about 295,000 reports, of which 241,744 were written by Italy-based evaluators.[23] Overall, 45% of reports express a positive vote (Appendix Table A7). Committee decisions display substantial unanimity: about 86% decisions are unanimous.

We use report text to construct proxies for evaluator effort. For each report, we measure length and lexical complexity. Complexity is based on an inverse document frequency (IDF) measure, which assigns greater weight to words that appear in fewer reports. After removing punctuation, numerals, stop-words, and first names, we stem words to account for gendered forms in Italian. We summarize complexity using *Average IDF*, the mean IDF-weighted rarity of words in the report,

22. As Bagues et al. (2019a) explain, in the first ASN round co-authors and colleagues were not formally subject to conflict-of-interest rules. Committees could self-impose restrictions, but only three committees required abstention, affecting 84 candidates who thus received only four reports.

23. Due to a technical issue, reports are missing for 202 applications.

and *Total IDF*, the sum of IDF weights. The average report contains 182 words; the mean *Average IDF* is 2.54, implying that the typical word appears in about 8% of reports. These measures differ substantially across disciplines: STEM reports are shorter and use less rare language than SSH reports, averaging 150 versus 233 words, with an *Average IDF* of 2.29 versus 2.94 and a *Total IDF* of 220 versus 438. STEM reports are also slightly more likely to mention bibliometric indicators, at 47% compared with 43% in SSH; all discipline differences are statistically significant at $p < 0.001$ (Appendix Table A7).

We validate these textual measures in two settings where greater evaluator effort is expected. Reports are longer and more complex when evaluators have stronger links to the candidate, such as prior co-authorship, even after conditioning on candidate and evaluator fixed effects (Appendix Table A8). Reports are also significantly longer and more lexically diverse for more productive candidates, whereas greater evaluator workload is associated with lower lexical diversity. Evaluators also write longer and more complex reports for marginal candidates whose qualification hinges on a single vote, conditional on candidate publication quantity and publication impact and evaluator fixed effects (Appendix Table A9). Together, these patterns provide support for interpreting textual characteristics as proxies for effort, while recognizing that they may omit other dimensions of effort devoted to evaluation.

4 Empirical analysis

This section first examines whether committees with stronger research records make qualification decisions that better predict candidates' future performance. We then turn to the costs of committee service for evaluators' subsequent research, supervision, and future service. Finally, we explore the mechanisms through which better-published members affect committee decisions.

4.1 Committee composition and selection outcomes

We assess decision quality using candidates' post-exam outcomes. Committees observe past publications and citations but must also evaluate less readily measured dimensions of potential, including the strength of candidates' research pipelines, the originality of their agendas, and the expected impact of ongoing work. If better-published committees interpret these signals more accurately, the candidates they qualify should perform better in the future.

This interpretation is complicated by the fact that committee quality may affect both screening and stringency. Better-published committees may rank candidates more accurately, but they

may also impose a higher qualification threshold. If stricter committees qualify fewer candidates, stronger outcomes among those qualified do not by themselves establish better screening.

We therefore proceed in two steps. First, we examine whether better-published committees are more selective and whether qualification outcomes relate differently to observable candidate characteristics. Second, we compare the subsequent outcomes of qualified and rejected candidates. If better-published committees merely raise the threshold, both groups should have stronger subsequent outcomes. If they screen more effectively, qualified candidates should perform better, whereas rejected candidates should not. Appendix Figure A1 illustrates this distinction.

Committee quality and the qualification decision. We begin by examining how committee composition affects qualification decisions:

$$Q_{i,e} = \beta_0 + \beta_1 T_e + \beta_2 N_i + \beta_3 I_i + \beta_4 (T_e \times N_i) + \beta_5 (T_e \times I_i) + \mathbf{X}'_{i,e}\Gamma + \varepsilon_{i,e} \tag{1}$$

where $Q_{i,e}$ is an indicator equal to one if candidate i qualifies in examination e. Committee quality, T_e, is defined as the average number of quality-adjusted publications of the committee members, normalized within the corresponding pool of eligible evaluators.[24] Candidate publication quantity, N_i, is the total number of publications in the previous ten years, and candidate publication impact, I_i, is the share of high-impact publications over the same period. Both variables are normalized among applicants to the same examination. The vector $\mathbf{X}_{i,e}$ includes four broad-discipline indicators (Sciences, Medicine and Veterinary, Engineering, and Social Sciences and Humanities) to absorb differences in qualification rates across broad fields. It also includes candidate–committee proximity controls – the number of evaluators from the same subfield, the same affiliation, and with prior coauthorship – to separate similarity in candidate–evaluator productivity profiles from professional proximity between candidates and evaluators.

The coefficient β_1 captures the effect of committee quality on overall selectivity, while β_4 and β_5 capture whether the relationship between qualification and candidates' publication quantity and impact varies with committee quality. Because T_e averages quality across the four Italy-based committee members, replacing an average evaluator with one whose pre-exam publication record is one standard deviation stronger increases T_e by one fourth. We report all estimates as the effect of replacing one average member with an evaluator whose publication record is one standard deviation

24. Results are robust to normalizing evaluator quality among all Full Professors in the corresponding discipline.

stronger.

Identification comes from the random assignment of evaluators to committees, conditional on the institutional constraints governing the draw. Committees are filled sequentially from a computer-generated random sequence, subject to restrictions on institutional representation and, in some fields, subfield quotas. We follow Borusyak and Hull (2023) by simulating the known assignment mechanism and constructing the residualized random component of the initially drawn committees' quality:

$$\widetilde{Z}_e = Z_e - \mathbb{E}[Z_e \mid \mathcal{W}_e]$$

where Z_e is the quality of the initially drawn committee and $\mathcal{W}_e$ denotes the features of the assignment mechanism: the eligible evaluator pool and the constraints imposed by the randomization procedure. We compute $\mathbb{E}[Z_e \mid \mathcal{W}_e]$ by simulating the draw algorithm one million times. The residualized variable, $\widetilde{Z}_e$, isolates the random variation generated by the conditionally random assignment mechanism. This instrument also accounts for potentially endogenous resignations following the initial random draw. We instrument T_e, $T_e \times N_i$, and $T_e \times I_i$ with $\widetilde{Z}_e$, $\widetilde{Z}_e \times N_i$, and $\widetilde{Z}_e \times I_i$, respectively. We estimate Equation (1) by two-stage least squares (2SLS) and cluster standard errors at the examination level. Predetermined candidate characteristics are balanced across the resulting variation in committee quality (Appendix Table A10), and, consistent with a low resignation rate, the first stage is strong with a Kleibergen–Paap rk Wald F-statistic of 128 (Appendix Table A11).

Table 1 reports the results. Better-published committees are more selective: replacing an average evaluator with one whose publication record is one standard deviation stronger reduces the qualification rate by about 1.5 percentage points, or 4 percent relative to the baseline rate of 37 percent (Column 1).[25]

Committee quality also changes how qualification outcomes relate to observable candidate characteristics. With better-published evaluators, qualification is less strongly associated with publication quantity and more strongly associated with publication impact (Column 2). Improving the quality of one member by one standard deviation reduces the association with publication quantity by about 5 percent and increases that with publication impact by about 9 percent. These patterns are similar in STEM+M and SSH fields (Columns 3 and 4).

This shift toward publication impact does not appear to reflect greater mechanical reliance on bibliometric measures. Reports written by better-published evaluators are, if anything, less likely to

25. A one-standard-deviation increase in average committee quality reduces the qualification probability by 0.059. Dividing by four gives the effect of replacing one committee member.

mention citations, the h-index, or publications in top journals: a one-standard-deviation increase in evaluator quality reduces the probability of mentioning any such indicator by 2.5 percentage points (Appendix Table A12).

Columns 5–8 instead examine promotion within five years. Committee quality has a small and statistically insignificant direct effect on the promotion rate. Thus, although better-published committees grant fewer qualifications, their greater selectivity does not reduce the number of eventual promotions. Instead, the composition of promoted candidates changes: promotion becomes more strongly associated with publication impact and less strongly associated with publication quantity. Again, the estimates are similar across STEM+M and SSH fields.

Overall, committee composition affects both qualification rates and the characteristics of candidates who advance. Better-published committees grant fewer qualifications, but this additional selectivity does not constrain university-level promotions. Rather, it shifts advancement toward candidates with greater research impact.

Committee quality and the quality of selection. We next examine whether the differences in qualification criteria documented above translate into better selection decisions. We estimate

$$y_{i,e,t+k} = \alpha^q + \beta^q T_e + \mathbf{X}_i' \mathbf{\Gamma}^q + u^q_{i,e,t+1} \quad \forall\, q \in \{0,1\} \tag{2}$$

separately for candidates who qualify ($q = 1$) and who are rejected ($q = 0$). The outcome variable $y_{i,e,t+k}$ measures candidate i's publications, citations, or career advancement over k years since the publication of the qualification outcomes in year t. Controls include four broad-discipline indicators, candidates' publication quantity, publication impact, citations, affiliation, and academic rank at the time of the exam. As before, we estimate this equation by 2SLS, instrumenting T_e with $\tilde{Z}_e$, and cluster standard errors at the examination level.

Conditioning on candidates' observable records allows us to test whether better-published committees identify dimensions of quality not readily captured in the data, such as the novelty or promise of a research agenda. This is the policy-relevant margin: observable criteria can be incorporated into formal rules, whereas expert evaluation is valuable insofar as it reveals less quantifiable dimensions of potential. Accordingly, β^1 measures whether better-published committees qualify candidates who subsequently outperform observationally similar peers, and β^0 provides the analogous comparison among rejected candidates.

A higher qualification threshold should improve the average residual quality of both qualified and

rejected candidates: those who pass are stronger, but so are those who narrowly fail. Better ranking instead implies stronger subsequent outcomes among qualified candidates without correspondingly stronger outcomes among those rejected. We therefore interpret

$$\beta^1 > 0 \qquad \text{and} \qquad \beta^0 \leq 0$$

as evidence consistent with improved selection rather than greater stringency alone.

This interpretation requires that qualification not have systematically different effects across candidates selected by committees of different quality. Appendix D shows that heterogeneous effects of qualification cannot explain our findings. It also requires candidate quality to be independent of assigned committee composition and committee composition to affect subsequent outcomes only through the qualification decision. Random assignment supports the first condition. The second is plausible because committees have no direct interaction with candidates and their reports justify decisions rather than provide individualized advice or assistance.

Table 2 reports the results. Committee quality has small and statistically insignificant effects on the subsequent publication output and number of high-impact articles of both qualified and rejected candidates (Columns 1–2). The pattern differs for citations (Column 3): qualified candidates evaluated by better-published committees receive more subsequent citations, whereas rejected candidates receive fewer. Replacing one average evaluator with one whose publication record is one standard deviation stronger raises qualified candidates' future citations by 1.1% of a standard deviation and lowers those of rejected candidates by 0.5% of a standard deviation. Equivalently, it increases the citation advantage of qualified candidates from 0.15 to 0.161 standard deviations, an improvement in selection performance of 7.3%.

The citation advantage appears for both pre- and post-exam publications but is larger for post-exam work (Columns 4–5). Better-published committees therefore appear not only to recognize the value of existing research more accurately, but also to identify candidates with stronger subsequent trajectories.

Career outcomes provide similar evidence (Column 6). Among Associate Professor applicants, better-published committees qualify candidates who are more likely to become Full Professors, without rejecting candidates who subsequently advance at higher rates. Replacing one committee member with an evaluator whose publication record is one standard deviation stronger raises a qualified candidate's probability of promotion to Full Professor within ten years by about 1 percentage

point, or 5.9% relative to the baseline.

We observe the same pattern for entry into the future ASN evaluator pool (Columns 7–8). Better-published committees qualify Associate Professor candidates who are more likely to serve as evaluators in future ASN cycles and more likely to do so as top evaluators. Replacing an average evaluator with one who is one standard deviation better published raises the probability that a qualified Associate Professor candidate later serves as an evaluator by 0.75 percentage points, or 7.5%, and the probability of becoming a top evaluator by 0.25 percentage points, or 6.25%.[26] Committee composition may therefore affect not only current selection but also the composition of future evaluator pools.[27]

As a complementary test, we compare candidates qualified by committees with above-mean evaluator-quality shocks with candidates unanimously qualified by committees below the mean. These groups account for nearly identical shares of their respective candidate pools – 34.1% and 33.8% – holding selectivity approximately fixed. Candidates selected by committees with positive quality shocks subsequently receive more citations and, among Associate Professor applicants, are more likely to become Full Professors (Appendix Table A14). This result provides a particularly direct test of ranking quality: even holding fixed the share of candidates selected, better-published committees choose candidates with stronger subsequent outcomes.

Taken together, the results are inconsistent with a uniform increase in the qualification threshold alone. Better-published committees select candidates with stronger subsequent citation and career outcomes, without rejecting candidates who later perform better. Across future citations, promotion to Full Professor, and entry into subsequent evaluator pools, replacing one average committee member with an evaluator whose publication record is one standard deviation stronger improves selection quality by approximately 6–7%.

4.2 Costs of committee service

We now use the random assignment of evaluators to estimate how committee service affects subsequent research, PhD supervision, and willingness to serve again. The analysis compares researchers assigned to committees with researchers who had the same *ex ante* assignment probabilities but were not drawn. We first estimate average effects on research output, then examine heterogeneity by discipline, evaluation workload, team size, and baseline researcher quality, and finally turn to

26. We classify individuals as *top researchers* at the time of the relevant evaluation cycle.

27. Estimates without controls for candidate observables are qualitatively similar but less precise; see Appendix Table Table A13.

supervision and future service.

Foregone research output. Using the universe of potential ASN evaluators between 2012 and 2021, we estimate:

$$y_{j,e,t} = \beta_0 + \beta_1 D_{j,e} + \mathbf{X}_j'\mathbf{\Gamma} + \mu_e + \varepsilon_{j,e,t} \tag{3}$$

where $y_{j,e,t}$ denotes the research output of researcher j in year t after examination e. We normalize research production among researchers in the same pool of potential evaluators. $D_{j,e}$ is an indicator variable equal to one if the researcher appears as an evaluator on an ASN committee. We initially use total publications – the broadest measure comparable across fields – as the outcome and examine its components below. The coefficient β_1 captures the causal effect of being assigned to committee service on subsequent research output.

We define

$$\tilde{Z}_{j,e} \equiv Z_{j,e} - \mathbb{E}[Z_{j,e}|\mathcal{W}_e]$$

where $Z_{j,e}$ is an indicator equal to one if the researcher is assigned to serve on the committee during the initial random draw, and $\mathbb{E}[D_{j,e} \mid \mathcal{W}_e]$ denotes the corresponding assignment probability under the known assignment mechanism. Thus, $\tilde{Z}_{j,e}$ captures the residual variation in committee assignment induced by the random draw, net of researchers' *ex ante* probability of selection.

We use this re-centered variable $\tilde{Z}_{j,e}$ as an instrument for whether the researcher serves as evaluator. To increase precision, we control for the researcher's pre-evaluation productivity trajectory, $\mathbf{X}_j$, including the number of publications, the share of high-impact articles, and a quadratic in academic age. We include examination fixed effects, μ_e, and normalize all productivity measures at the examination level to account for field-specific differences. We estimate Equation (3) by 2SLS and cluster standard errors at the examination level. Predetermined evaluator characteristics are balanced across the instrument (Appendix Table A15), and the first stage is strong, with a Kleibergen–Paap rk Wald F-statistic above 16,000 (Appendix Table A16).

Panel A of Figure 2 shows that committee service reduces total publications over the following three years by approximately 5% of a standard deviation. This corresponds to about 1.5 publications, or 20% of an average evaluator's annual output.[28]

Panel B reports cumulative effects, normalized to zero in the year preceding the examination.[29]

28. The standard deviation of annual output is approximately 10 publications, while mean annual output is 6.5; see Appendix Table A3.

29. For pre-exam years, the outcome is defined symmetrically as cumulative output between that year and the year

The decline remains statistically significant for five years. Although the estimates attenuate thereafter, they remain negative at longer horizons, suggesting a persistent disruption rather than a short delay in publication.

These estimates likely provide a lower bound on the cost of service. They apply to researchers who volunteer and whose participation responds to random assignment. Researchers expecting especially high costs may not volunteer, while those who incur large costs after being drawn may be more likely to resign. Both forms of selection would attenuate the estimated loss relative to that among all eligible researchers.

Heterogeneity by field, evaluation workload, and team size. The average effect may conceal substantial heterogeneity across disciplines. Evaluation may require different amounts of effort across fields, as reflected in large differences in the length and detail of reports. Research production may also differ in its capacity to absorb time shocks: STEM+M researchers typically work in larger teams, which may redistribute tasks when one member is diverted to service.

Table 3 shows that committee service has a small and statistically insignificant effect on publication output in STEM+M, but reduces output in SSH by approximately 10% of a standard deviation over three years.

To distinguish possible mechanisms, we examine variation within each broad field in expected evaluation workload and baseline team size. We measure workload by the discipline-level average number of words written across evaluators' individual reports in 2012, capturing both the volume of candidates and the effort devoted to their evaluation. Team size is the researcher's average number of authors per paper before the examination.

We find no significant treatment-effect differences by expected workload within either STEM+M or SSH. By contrast, the SSH effect is concentrated among researchers working in smaller teams. Effects in STEM+M are smaller and show no clear team-size gradient. Team-size distributions, however, overlap little across fields: even high-collaboration SSH researchers work in smaller teams than low-collaboration STEM+M researchers. The evidence therefore suggests that collaboration networks help absorb shocks to researchers' time. This need not reduce the total cost of service: in larger teams, some of the burden may fall on collaborators rather than appearing entirely in the evaluator's own output.

Motivated by these patterns, we adopt the following specification in which the treatment effect

preceding the examination.

varies continuously with baseline team size:

$$y_{j,e,t} = \beta_0 + \beta_1 \frac{D_{j,e}}{\bar{\alpha}_{j,e}} + \beta_2 \frac{1}{\bar{\alpha}_{j,e}} + \mathbf{X}_j'\mathbf{\Gamma} + \mu_e + \varepsilon_{j,e,t} \tag{4}$$

where $\bar{\alpha}_{j,e}$ denotes researcher j's average number of authors per paper during the ten years preceding the draw for exam e. We instrument the scaled treatment variable, $D_{j,e}/\bar{\alpha}_{j,e}$, with the correspondingly scaled initial random draw, $\widetilde{Z}_{j,e}/\bar{\alpha}_{j,e}$.

This specification parsimoniously captures the treatment-effect heterogeneity associated with collaboration intensity and discipline, improves precision, and avoids arbitrary sample splits. We do not interpret it as identifying the causal effect of additional collaborators. A horse-race specification fits the data better than one containing only the unscaled treatment indicator (Appendix Table A17).[30] Under this formulation, β_1 is the effect for a researcher who sole-authors all papers; for the average researcher, whose papers have 5.93 authors, the implied effect is $\beta_1/5.93$.

Figure 3 reports the cumulative estimates. For a sole author, committee service reduces output over three years by approximately 19% of a standard deviation, with statistically significant effects lasting up to eight years. Sole authors constitute less than 5% of the sample. At the sample mean of 5.93 authors per paper, the implied loss is approximately 3% of a standard deviation, or 1.2 publications. In economics, the corresponding losses are 1.9 publications for a sole author and 0.6 for a researcher with the field-average team size of three.

We use this scaled specification below to examine heterogeneity by baseline researcher quality while accounting for systematic variation with team size.

Heterogeneity by evaluator quality. We next test whether the opportunity cost of service rises with evaluators' baseline research quality. We interact the scaled treatment with both continuous quality-adjusted publication output and an indicator for *top researchers*.

Table 4 reports estimates over three- and eight-year horizons. Higher-quality researchers experience larger productivity losses, especially in the short run. For a sole author of average quality, committee service reduces three-year output by approximately 19% of a standard deviation. The estimated loss is about 1.6 times larger for a researcher whose publication record is one standard deviation stronger.

Panels A and B of Figure 4 plot implied effects across the team-size distribution for top and

30. The results are robust to scaling by the inverse square and inverse square root of team size.

other researchers. At the average team size, committee service reduces short-run output by 5.7% of a standard deviation for top researchers and 1.6% for other evaluators, corresponding to approximately 2.2 and 0.6 publications, respectively.[31]

Long-run heterogeneity is less precisely estimated, but the losses appear more persistent for top researchers. Eight years after assignment, a top researcher working in a team of average size has produced 3.7% of a standard deviation fewer publications, a cumulative loss of approximately 1.9 publications. The largest effects occur among top researchers working alone. Committee service reduces their three-year output by as much as 33% of a standard deviation, compared with approximately 10% for other sole authors. These losses also represent a larger share of top researchers' usual production, consistent with both a higher opportunity cost of their time and less scope for adjustment through collaborators.

Appendix Figure A2 examines which outputs decline. The reduction in journal articles is large and statistically significant among top researchers. Estimates are negative for all outcomes, but larger and more precise for lower-impact articles and other research outputs than for high-impact articles or citations.

Overall, committee service imposes larger and more persistent productivity costs on better-published researchers, particularly those working in small teams. The results are consistent with collaboration networks helping to absorb temporary demands on researchers' time.

Foregone student supervision. We next examine whether committee service affects PhD supervision. We estimate Equation (4) with the number of supervised PhD graduates as the outcome, restricting attention to universities that systematically report supervision activity.

Figure 5 shows that committee service reduces PhD supervision. The effects become statistically significant four years after the exam, consistent with reduced intake of new PhD students rather than delayed completion of ongoing supervision. In the long run, an average evaluator who serves on a committee supervises approximately 0.4 fewer PhD graduates.

Unlike research output, we find no meaningful heterogeneity in supervision losses by researcher quality (Appendix Figure A3). Consistently, descriptive statistics in Appendix Tables A2 and A3 show that the number of PhD graduates supervised is broadly similar across researchers of different

31. In the continuous-quality specification, at the average team size a one-standard-deviation increase in quality implies an additional short-run loss of approximately 2.3% of a standard deviation (0.136/5.93), relative to a baseline loss of 3.3% (0.193/5.93). Given the functional form, the statistical significance of this difference does not vary with team size.

productivity levels.

Future participation of selected researchers. Finally, we examine whether the costs of committee service affect evaluators' willingness to volunteer again. As discussed in Section 2.2, top researchers were initially more likely to volunteer, but their participation declined over time.

Table 5 show results from estimation Equation (4) using as an outcome an indicator for volunteering in the next cycle in which all members of the relevant evaluator pool are legally eligible. For the 2012 cohort, this is the 2018 cycle; for the 2016 cohort, it is 2023. Evaluators first drawn in 2018 or 2021 are excluded because no subsequent eligible cycle is observed.

A key challenge in this analysis is that participation rules create a mechanical relationship between initial assignment and later eligibility. Evaluators not drawn in the initial cycle may serve in an intermediate cycle and thereby become ineligible for the later cycle used as the outcome. Estimates from the unrestricted sample therefore conflate behavioral responses with this rule-based channel; accordingly, the positive but imprecise estimate in Column 1 is difficult to interpret.

To remove this mechanical effect, we exclude evaluators who were not initially drawn but subsequently served in an intermediate cycle – 2016 for the 2012 cohort and 2018–2021 for the 2016 cohort. This restriction may induce some selection because researchers more willing to volunteer are more likely to serve in intermediate cycles, but it eliminates the direct eligibility distortion.

Columns 2 and 3 show that prior service significantly reduces subsequent volunteering, especially among more productive researchers. For a sole author of average productivity, assignment to committee service lowers the probability of volunteering again by 8.4 percentage points. A one-standard-deviation increase in baseline productivity adds a further 6.9-percentage-point decline. Evaluated at the sample-average team size, the corresponding reductions are 1.4 percentage points for a researcher of average productivity and 2.6 percentage points for one whose productivity is one standard deviation higher.

Thus, committee service discourages future participation, particularly among the most productive researchers. Combined with the larger research costs they incur, these results point to endogenous attrition from the evaluator pool among those whose expertise is most valuable.

4.3 Mechanisms of Expert Influence

Better-published evaluators may improve decisions through their own assessments or by changing how their colleagues evaluate candidates. We distinguish these channels using evaluator–candidate-

level votes and individual report characteristics.

Experts' influence on voting. To assess whether the better-published researchers apply different criteria and whether their presence affects other evaluators' behavior, we estimate the following equation:

$$y_{i,j,e} = \beta_0 + \beta_1 T_j + \beta_2 T_{e,-j} + \mathbf{X}'_{i,j}\mathbf{\Gamma} + \varepsilon_{i,j,e} \tag{5}$$

where the outcome variable is a binary measure indicating whether the evaluator j casted a positive vote for candidate i in examination e. The vector $\mathbf{X}_{i,j}$ includes candidate productivity measures and evaluator–candidate connections, such as prior co-authorship, belonging to the same subfield, and being affiliated with the same university. The term T_j denotes the quality of evaluator j, and the term $T_{e,-j}$ denotes the average quality of all other evaluators on the committee, excluding evaluator j, defined as:

$$T_{e,-j} = \frac{1}{N_e - 1} \sum_{\{k \neq j\} \in e} drawn_{k,e} \times T_k$$

with $drawn_{k,e}$ indicating whether evaluator k is selected for committee e, and N_e denoting committee size. Similarly to earlier sections, we follow the logic of Borusyak and Hull (2023) to isolate random variation in peer quality. Specifically we instrument the quality of realized peers $T_{e,-j}$ with the residual random component that arises due to the randomness of the draw:

$$\tilde{Z}_{e,-j} = Z_{e,-j} - \mathbb{E}[Z_{e,-j}|\mathcal{W}_e]$$

where $Z_{e,-j}$ is the quality of peers following the initial random draw, prior to any potential resignations and $\mathbb{E}\left[Z_{e,-j}|\mathcal{W}_e\right]$ captures the expected peer quality faced by evaluator j, conditional on them being drawn and the institutional constraints of the conditional random draw procedure:

$$\mathbb{E}\left[Z_{e,-j}\right] = \frac{1}{N_e - 1} \sum_{\{k \neq j\} \in e} \mathbb{P}\left[drawn_{k,e} \mid \mathcal{W}_e, drawn_{j,e} = 1\right] \times T_k$$

where $\mathbb{P}\left[drawn_{k,e} \mid \mathcal{W}_e, drawn_{j,e} = 1\right]$ denotes the conditional probability that evaluator k is selected for committee e, given that evaluator j is also selected. These probabilities are computed as averages over one million simulated draws that replicate the official committee selection procedure.

In this specification, β_1 captures the effect of an evaluator's own quality on the probability

of casting a positive vote, holding the quality of their peers fixed. The coefficient β_2 reflects the impact of increasing the quality of peers on the committee. Since committees consist of four Italian members, $\beta_2/3$ gives the effect of replacing one average peer with an evaluator whose publication record is one standard deviation stronger, holding the quality of the other peers fixed.

We estimate Equation (5) by 2SLS, instrumenting $T_{e,-j}$ with $\tilde{Z}_{e,-j}$, and cluster standard errors at the examination level.

Table 6 reports the results. Better-published evaluators are individually stricter: a one-standard-deviation increase in own quality is associated with a 2.3-percentage-point lower probability of voting positively (Column 1). Randomly assigned peer quality also affects voting. Replacing one average peer with an evaluator whose publication record is one standard deviation stronger reduces an evaluator's probability of voting positively by approximately 1.8 percentage points.

Column 2 compares evaluators voting on the same candidate. Within candidates, evaluator quality is not associated with a differential propensity to vote positively. Together with Column 1, this pattern suggests that better-published peers shift committee-wide standards and induce convergence in voting rather than simply casting more negative votes themselves.

Columns 3 and 4 examine how votes relate to candidate characteristics. Better-published evaluators' votes are less strongly associated with publication quantity and more strongly associated with publication impact. The randomly assigned quality of their peers produces a similar shift: replacing one peer with an evaluator whose publication record is one standard deviation stronger reduces the marginal association with publication quantity by about 0.6 percentage points and increases that with publication impact by about 1 percentage point.

Thus, better-published evaluators influence decisions both through their own assessments and through their colleagues. The peer effect on voting is nearly as large as the association with an evaluator's own quality.

Experts' influence on effort. We next estimate Equation (**??**) using features of evaluator j's report on candidate i as proxies for effort: report length, *Average IDF*, which measures the rarity of the words used relative to other reports, and *Total IDF*, the sum of rarity-weighted words.

Table 7 shows no systematic relationship between an evaluator's own quality and any of these measures. Better-published evaluators neither write shorter reports nor use systematically different or more specialized language.

By contrast, randomly assigned peer quality increases all three effort proxies. Replacing one

average peer with an evaluator whose publication record is one standard deviation stronger raises report effort by approximately 6–7% of a standard deviation, depending on the measure.

Better-published evaluators therefore influence not only the criteria applied by their colleagues but also the effort those colleagues exert. This pattern is consistent with the reputational mechanism in Swank and Visser (2023): evaluators may exert greater effort when serving alongside more accomplished researchers because they care about their reputation in the eyes of those colleagues.

Non-linear impact of *top researchers* on committees. The peer effects above imply that the value of expert presence may be nonlinear. If experts affected decisions only through their independent votes, their influence should increase approximately proportionally with their number. If they also shift common standards or induce greater effort, the first expert may influence several non-expert colleagues, generating diminishing marginal effects.

We test this possibility by replacing continuous committee quality with indicators for the number of *top researchers* and controlling for the simulated probability to have one, two, or three or more top researchers on the committee. The presence of at least one top researcher lowers the qualification probability by 5.8 percentage points (Appendix Table A18). Relative to committees with none, committees with one, two, and at least three top researchers are respectively 4.2, 7.3, and 8.6 percentage points less likely to qualify a candidate. Strictness therefore rises with the number of top researchers, but less than proportionally. Although the confidence intervals do not sharply distinguish every adjacent category, the pattern is consistent with diminishing marginal effects.

Effort responses exhibit similar nonlinearity (Appendix Table A19). Relative to evaluators with no top-researcher peers, those with one such peer write reports that are 30% of a standard deviation longer. The effect is somewhat larger with two top-researcher peers but does not increase further with three or more.[32]

The estimates for selected candidates' future outcomes do not reveal a clear pattern of diminishing returns across measures: they are too imprecise to distinguish this pattern from alternative relationships (Appendix Table A20).

Overall, top researchers affect committees through more than their own votes. Their presence changes both the standards applied and the effort exerted by other evaluators, with the largest

32. We find no statistically significant differences in evaluators' productivity losses by the quality of their randomly assigned peers (Appendix Table A21, Panel A). The lower bound of the 95% confidence interval nevertheless allows an additional loss of up to 2.6% of a standard deviation, relative to the 3% average loss associated with service for a researcher with the mean team size.

incremental effect arising when expert representation is initially low.

5 Conclusion

This paper provides causal evidence on the benefits and costs of assigning better-published researchers to collective evaluation. Using the random assignment of evaluators to Italy's national academic qualification committees, we show that better-published researchers improve selection, influence their colleagues' behavior, and incur substantial opportunity costs when assigned to serve.

Better-published committees are stricter, but their contribution extends beyond raising the qualification threshold. Their decisions are less strongly associated with publication quantity and more strongly associated with publication impact. Conditional on observable past productivity, the candidates they qualify have stronger subsequent citation and career outcomes, whereas those they reject do not subsequently perform better. Replacing an average committee member with an evaluator whose publication record is one standard deviation stronger improves selection by approximately 6–7%, depending on the ex-post outcome considered. Expert judgment therefore reveals dimensions of candidate potential not captured by observable records, even in a setting with extensive bibliometric information.

Expertise also operates through peer influence. Better-published evaluators are themselves stricter and apply different criteria, but their presence changes how other members evaluate candidates. Evaluators randomly assigned to more accomplished peers adopt similar criteria and write longer, more detailed reports, consistent with within-committee reputational pressures.

These benefits come with persistent costs that extend beyond the evaluation itself. Committee service reduces evaluators' publication output for several years, with cumulative losses equivalent to about 20% of a typical year's production, and also lowers subsequent PhD supervision. The research losses are largest among better-published researchers and among those working in small teams, whose collaborators have less scope to absorb disruptions to their time. Larger teams appear better able to redistribute work, although this may shift rather than eliminate the cost by imposing additional burdens on coauthors. Evaluation service can therefore affect not only evaluators, but also students, collaborators, as well as readers and policymakers whose access to new knowledge is delayed.

Service also affects the future supply of evaluation. Researchers randomly assigned to committees become less likely to volunteer again, with larger declines among more productive researchers. At the same time, better selection replenishes the future evaluator pool by advancing candidates

who are more likely to become qualified evaluators themselves. Evaluation systems thus generate opposing dynamic forces: expert participation improves the future pool of researchers and evaluators, but the costs of participation induce valuable experts to withdraw. The scarcity of qualified evaluators is consequently partly shaped by how evaluation systems allocate and reward expert time.

These findings shift the relevant design question from whether expert judgment is valuable to where its value exceeds its social cost. Assigning the strongest available researchers to every committee may improve current decisions, but it also concentrates burdens on researchers whose alternative contributions are especially valuable and may reduce their future participation. Systems that rely heavily on expert judgment without addressing its costs may gradually lose precisely the evaluators who contribute most to decision quality. More generally, committee design should account for peer effects, heterogeneous opportunity costs, and future evaluator supply rather than maximize current evaluator quality in isolation.

Voluntary participation alone need not produce an efficient allocation, because evaluators may not internalize costs borne by students, collaborators, other institutions, or the future scientific community. Reducing unnecessary evaluation demands, distributing service more broadly, and compensating or protecting the research and mentoring time of heavily solicited experts may be necessary to help sustain peer evaluation.

Although our analysis focuses on academic promotion committees, the forces we document are not specific to academia. Grant panels, journal editorial boards, regulatory agencies, and professional licensing bodies also rely on repeated participation by highly skilled professionals whose time is valuable elsewhere. In such settings, decision quality depends not only on the expertise of individual members, but also on how expertise shapes peer behavior and on the opportunity costs borne by those who generate the largest collective benefits. The central challenge is not to replace expert judgment with mechanical indicators, but to design institutions that preserve the scarce expertise on which credible judgment depends.

References

Aczel, B., B. Szaszi, and A. O. Holcombe (2021): "A Billion-Dollar Donation: Estimating the Cost of Researchers' Time Spent on Peer Review," *Research Integrity and Peer Review*, 6, 14.

Arts, S., N. Melluso, and R. Veugelers (2025): "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, 527–554.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017): "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review*, 107, 1207–38.

——— (2019a): "Connections in Scientific Committees and Applicants' Self-Selection: Evidence from a Natural Randomized Experiment," *Labour Economics*, 58, 81–97.

——— (2019b): "A Walk on the Wild Side: 'Predatory' Journals and Information Asymmetries in Scientific Evaluations," *Research Policy*, 48, 462–477.

Bertocchi, G., A. Gambardella, T. Jappelli, C. A. Nappi, and F. Peracchi (2015): "Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy," *Research Policy*, 44, 451–466.

Borusyak, K. and P. Hull (2023): "Nonrandom Exposure to Exogenous Shocks," *Econometrica*, 91, 2155–2185.

Brogaard, J., J. Engelberg, and C. A. Parsons (2014): "Networks and Productivity: Causal Evidence from Editor Rotations," *Journal of Financial Economics*, 111, 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, 269–327.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

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

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

Dance, A. (2023): "Stop the Peer-Review Treadmill. I Want to Get Off," *Nature*, 614, 581–583.

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

DORA (2012): "San Francisco Declaration on Research Assessment," https://sfdora.org/read/.

ERC (2026): "Applying for an ERC Grant in the 2027 Competitions: What You Need to Know," https://erc.europa.eu/news-events/news/applying-erc-grant-2027-competitions-what-you-need-know.

Funk, P., N. Iriberri, and N. Venus (2024): "Women in Editorial Boards: An Investigation of Female Representation in Top Economic Journals," *CEPR Discussion Papers*.

Hager, S., C. Schwarz, and F. Waldinger (2024): "Measuring Science: Performance Metrics and the Allocation of Talent," *American Economic Review*, 114, 4052–4090.

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

Heckman, J. J. and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," *Handbook of econometrics*, 6, 4875–5143.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols (2015): "Bibliometrics: The Leiden Manifesto for Research Metrics," *Nature*, 520, 429–431.

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

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

Italian National University Council (2023): "Proposta Su Abilitazione Scientifica Nazionale," https://www.cun.it/provvedimenti/sessione/329/analisi_e_proposte/analisi-proposta-del-20-aprile-2023.

Kolarz, P., A. Vingre, A. Vinnik, A. Neto, C. Vergara, C. O. Rodriguez, K. Nielsen, and L. Sutinen (2023): "Review of Peer Review," Tech. rep., Technopolis Group.

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

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

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

Merton, R. K. (1957): "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *American Sociological Review*, 22, 635–659.

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

Nieddu, M. and L. Pandolfi (2022): "The Effectiveness of Promotion Incentives for Public Employees: Evidence from Italian Academia," *Economic Policy*, 37, 697–748.

Ottaviani, M. and P. Sørensen (2001): "Information Aggregation in Debate: Who Should Speak First?" *Journal of Public Economics*, 81, 393–421.

Prat, A. (2005): "The Wrong Kind of Transparency," *American Economic Review*, 95, 862–877.

Priem, J., H. Piwowar, and R. Orr (2022): "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," .

Publons (2018): "Publons' Global State Of Peer Review 2018," Tech. rep., Publons, London, UK.

ROARS (2014): "Sulla revisione dell'ASN: alcune proposte," https://www.roars.it/sulla-revisione-dellasn-alcune-proposte/.

Stephan, P. (2010): "The Economics of Science," Handbook of the Economics of Innovation, Elsevier.

Stephan, P., R. Veugelers, and J. Wang (2017): "Reviewers Are Blinkered by Bibliometrics," *Nature*, 544, 411–412.

Swank, O. H. and B. Visser (2023): "Committees as Active Audiences: Reputation Concerns and Information Acquisition," *Journal of Public Economics*, 221, 104875.

Vesper, I. (2018): "Peer Reviewers Unmasked: Largest Global Survey Reveals Trends," *Nature*.

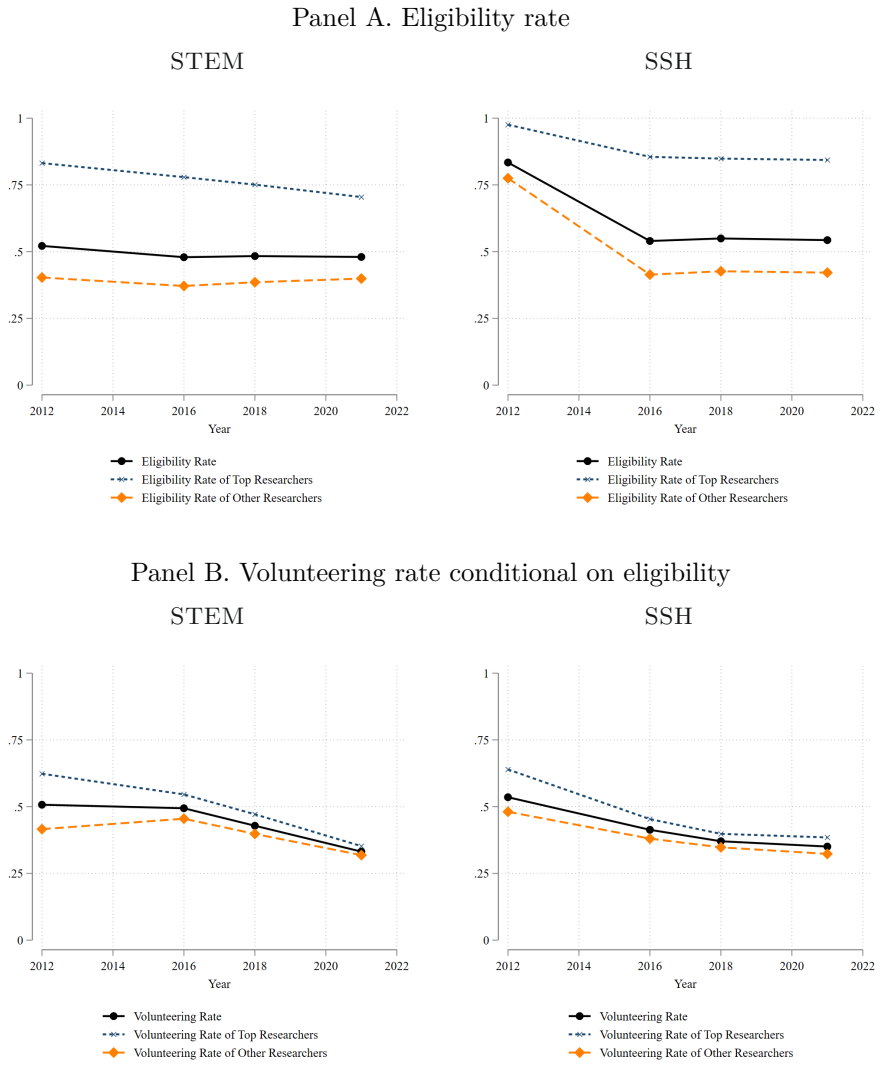
Visser, B. and O. H. Swank (2007): "On Committees of Experts," *The Quarterly Journal of Economics*, 122, 337–372.

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

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

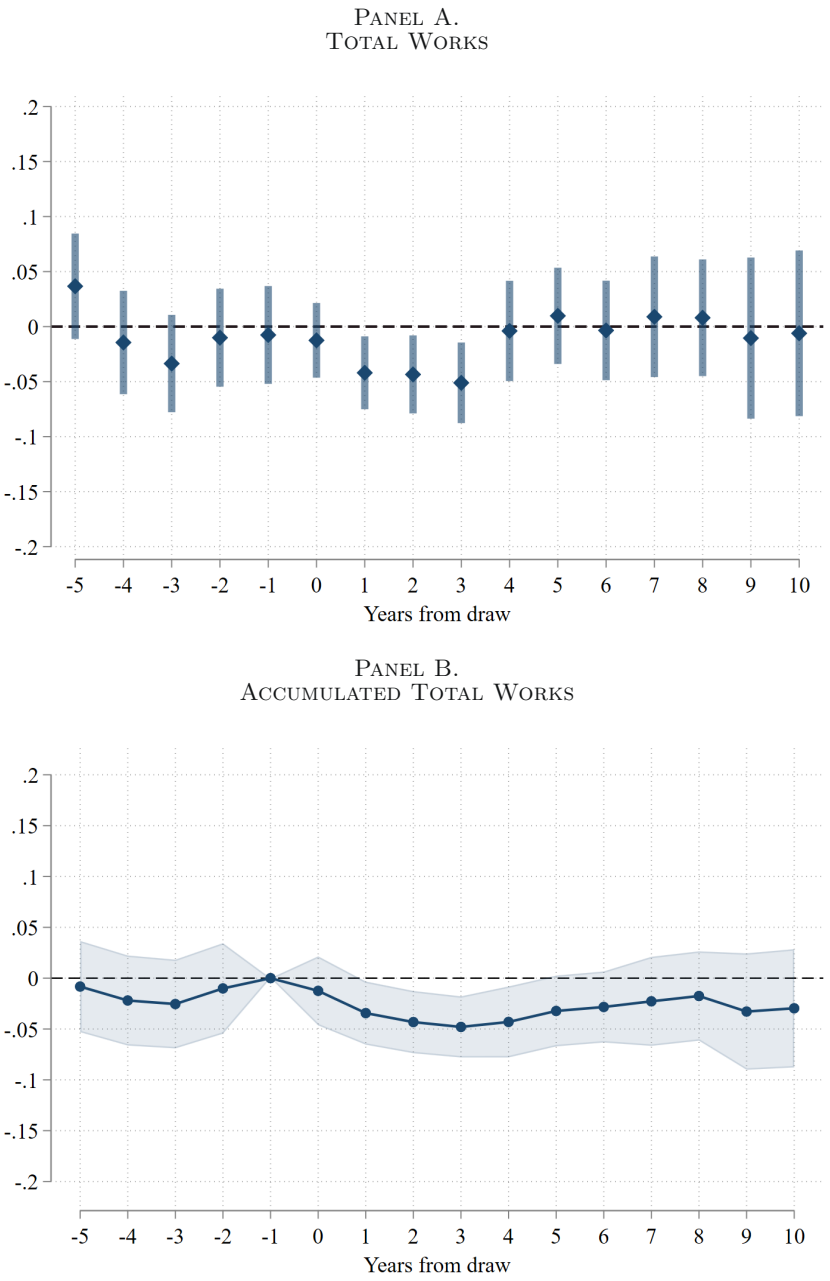
Figures

Figure 1: Eligibility and volunteering rate



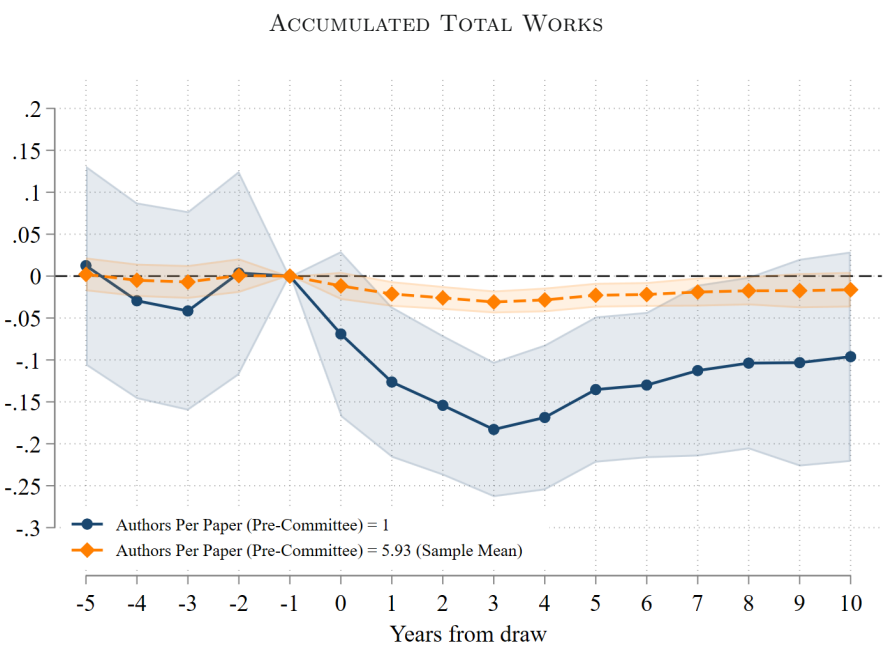
Notes: In Panel A. the sample is Full Professors at public universities over the period 2012-2021. In Panel B. the sample is all researchers at public universities who were eligible to participate as evaluators in the Italian evaluation system over the same period. We simulate eligibility following the official criteria published by the Ministry. Panel A. reports the share of eligible researchers who were observed in the pool of potential evaluators. Panel B. reports the share of eligible evaluators who volunteered. The solid black line shows trends for all researchers. The short-dashed blue line shows trends for top researchers. The long-dashed orange line shows trends for other researchers. *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.

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



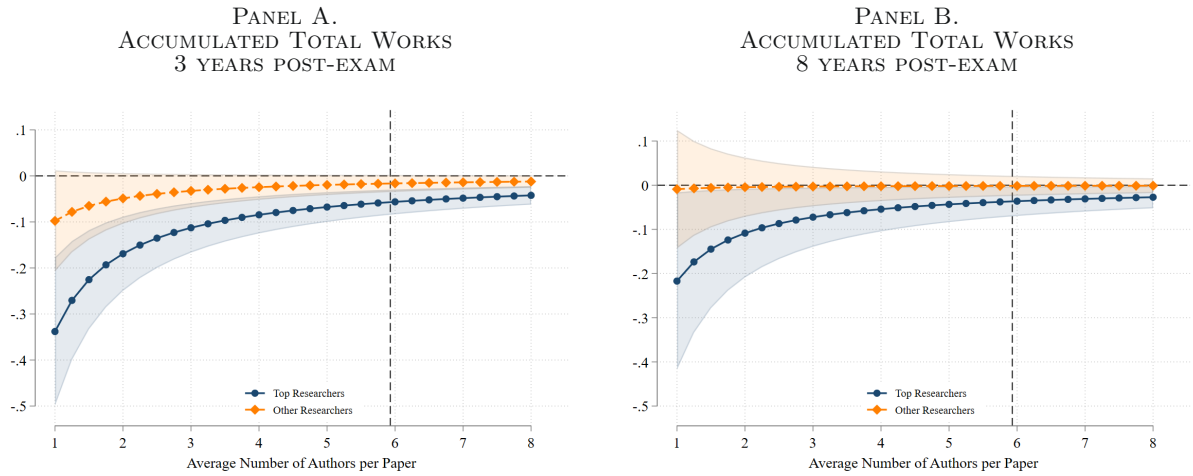
Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. The estimates are based on Equation (3). In Panel A. the outcome variable is total research output in a given year, normalized at the field-year level. In Panel B. the outcome variable is accumulated total research output starting from the year prior to the exam, normalized at the field-year level. We control for the probability of being drawn, and exam fixed effects. For years after the exam, we also control for past research production, and a second-order polynomial in academic age. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level. The vertical bars and shaded area denote the 95% confidence interval around the estimates.

Figure 3: Impact of sitting on the committee on total research output by the number of collaborators per paper prior to the exam



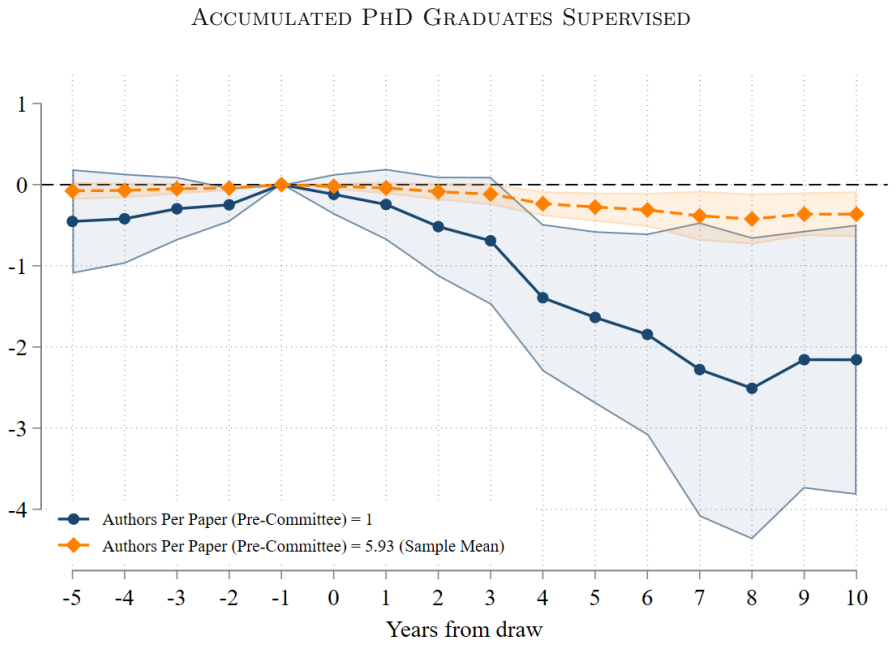
Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. The estimates are based on Equation (4). The outcome variable is accumulated total research output starting from the year prior to the exam, normalized at the field-year level. The solid blue line shows estimates for researchers who exclusively solo-authored in the ten years preceding the exam. The orange dashed line shows the estimates for the average researcher in our sample. We control for the probability of being drawn, and exam fixed effects. For years after the exam, we also control for past research production, and a second-order polynomial in academic age. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level. The shaded areas denote the 95% confidence interval around the estimates.

Figure 4: Impact of sitting on the committee on total research output by top researcher over the distribution of collaborators per paper prior to the exam



Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. The estimate are based on Equation (4). In Panel A. the outcome variable is accumulated total research output in the first three years following the exam. In Panel B. the outcome variable is accumulated total research output in the first eight years following the exam. The dashed line on the x-axis researchers' average number of authors per paper in the sample (5.93). The solid blue line shows estimates for top researchers. The orange dashed line shows estimates for other researchers. *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. We control for the probability of being drawn, exam fixed effects, past research production, and a second-order polynomial in academic age. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level. The shaded areas denote the 95% confidence interval around the estimates.

Figure 5: Impact of sitting on the committee on total number of PhD graduates supervised by the number of collaborators per paper prior to the exam



Notes: The sample is researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021, and whose universities consistently reported PhD students over the sample period. The estimates are based on Equation (4). The outcome variable is the accumulated number of PhD students graduating under researchers' supervision starting from the year prior to the exam, normalized at the field-year level. The solid blue line shows estimates for researchers who exclusively solo-authored in the ten years preceding the exam. The orange dashed line shows the estimates for the average researcher in our sample. We control for the probability of being drawn, and exam fixed effects. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level. The shaded area denotes the 95% confidence interval around the estimates.

Tables

Table 1: Committee composition and evaluation outcomes

	Qualified				Promoted			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			STEMM	SSH			STEMM	SSH
Candidate's Number of Publications	0.118***	0.118***	0.126***	0.106***	0.087***	0.087***	0.090***	0.083***
	(0.005)	(0.004)	(0.006)	(0.006)	(0.003)	(0.003)	(0.005)	(0.004)
Candidate's Impact of Publications	0.094***	0.094***	0.090***	0.097***	0.039***	0.039***	0.031***	0.051***
	(0.006)	(0.005)	(0.007)	(0.008)	(0.003)	(0.002)	(0.003)	(0.004)
Committee Quality	-0.059**	-0.059**	-0.061**	-0.052	0.011	0.011	0.016	-0.001
	(0.029)	(0.029)	(0.027)	(0.070)	(0.014)	(0.014)	(0.017)	(0.024)
Candidate's Number of Publications × Committee Quality		-0.025**	-0.026	-0.019		-0.007	-0.009	0.002
		(0.012)	(0.017)	(0.016)		(0.007)	(0.009)	(0.013)
Candidate's Impact of Publications × Committee Quality		0.034***	0.026*	0.051**		0.020***	0.015*	0.024**
		(0.012)	(0.015)	(0.019)		(0.007)	(0.009)	(0.011)
Observations	69020	69020	41395	27625	69020	69020	41395	27625
Adjusted R^2	0.117	0.118	0.120	0.115	0.054	0.055	0.059	0.052
Exams	184	184	109	75	184	184	109	75

Notes: The table reports estimates of Equation (1). The outcome is an indicator for qualification in Columns 1–4 and for promotion to the rank sought within five years in Columns 5–8. The sample includes all candidates who applied in the first wave of the 2012 cycle of the Italian evaluation system. *Candidate's Number of Publications* (N_i) measures total research output. *Candidate's Impact of Publications* (I_i) is the share of the candidate's articles published in top-quartile Web of Science journals in STEM+M or A-list journals in SSH. *Committee Quality* is the average quality-adjusted publication record of committee members during the ten years preceding the exam. Realized committee quality is instrumented by its deviation from the value expected under the assignment mechanism. Candidate productivity measures are standardized within exams, and evaluator quality is standardized within eligibility pools. All specifications include indicators for broad disciplinary groups and controls for evaluator–candidate connections through coauthorship, common institutional affiliation, and subfield. Standard errors are clustered at the exam level. *** $p < .01$, ** $p < .05$, * $p < .1$.

Table 2: Committee composition and the predictive content of qualification decisions

	All Candidates					Associate Professor Candidates		
	(1) Total Publications	(2) High-Impact Articles	(3) Citations (All Articles)	(4) Citations (Pre-Exam Articles)	(5) Citations (Post-Exam Articles)	(6) Promoted Full Prof.	(7) Future Evaluator	(8) Future Top Evaluator
Panel A. Successful Candidates								
Committee Quality	0.002	0.009	0.044***	0.031**	0.041***	0.038***	0.030***	0.010**
	(0.019)	(0.020)	(0.014)	(0.012)	(0.015)	(0.014)	(0.009)	(0.004)
Observations	25337	25337	25337	25337	25337	17536	17536	17536
R^2	0.19	0.16	0.37	0.61	0.16	0.02	0.01	0.02
Baseline	0.25	0.27	0.15	0.19	0.11	0.17	0.10	0.04
Panel B. Unsuccessful Candidates								
Committee Quality	-0.008	-0.006	-0.020**	-0.011	-0.018*	0.005	-0.000	0.000
	(0.014)	(0.017)	(0.010)	(0.008)	(0.010)	(0.005)	(0.003)	(0.002)
Observations	43676	43676	43676	43676	43676	29882	29882	29882
R^2	0.17	0.15	0.35	0.62	0.15	0.02	0.01	0.00
Baseline	-0.14	-0.16	-0.09	-0.11	-0.06	0.04	0.02	0.01

Notes: The table reports estimates of Equation (2). The dependent variables measure candidate outcomes ten years after the evaluation. In columns (1)–(5) the sample is all candidates who submitted an application in the first wave of the 2012 cycle of the Italian evaluation system. In columns (6)–(8) we restrict the sample to candidates to Associate Professor. *Promoted Full Prof.* is a dummy variable indicating whether the candidate is promoted to Full Professor within ten years. *Future Evaluator* is a dummy variable indicating whether the candidate appears in the pool of potential evaluators in any future cycle. Similarly, *Future Top Evaluator* is equal to one if the candidate appears in the pool of evaluators and is also coded as a top researcher. *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. *Committee Quality* is measured as the average quality-adjusted publication record of evaluators over the ten years preceding the exam. The final composition of the committee is instrumented by the deviation of realized committee quality from its expected value under the assignment mechanism. Both committee quality and candidate outcome variables are normalized at the examination level. We include four dummy variables for broad disciplinary group, control for candidates' number and impact of publications, and the number of pre-exam citations. We also include indicators for candidates' affiliation and academic rank at the time of the evaluation. Standard errors are clustered at the examination level.

*** $p < .01$, ** $p < .05$, * $p < .1$

Table 3: Impact of sitting on the committee on total research output by average by discipline, reports length, and collaborator team size

Panel A: STEMM					
	Overall	Evaluator Report Length		Collaboration Team Size	
	(1)	(2)	(3)	(4)	(5)
		Low	High	Low	High
Drawn Researcher	-0.024	-0.020	-0.038	-0.031	-0.018
	(0.018)	(0.027)	(0.026)	(0.029)	(0.025)
Observations	9775	4349	5248	4878	4834
Adjusted R^2	0.64	0.64	0.63	0.59	0.69
Average Number of Authors per Paper	7.96	7.43	8.33	4.37	11.60
Average Number of Words	59131	23573	88597	63083	55229
Panel B: SSH					
	Overall	Evaluator Report Length		Collaboration Team Size	
	(1)	(2)	(3)	(4)	(5)
		Low	High	Low	High
Drawn Researcher	-0.101***	-0.110***	-0.080**	-0.181***	-0.045
	(0.027)	(0.040)	(0.037)	(0.040)	(0.041)
Observations	5916	2611	3281	2921	2988
Adjusted R^2	0.48	0.50	0.48	0.50	0.48
Average Number of Authors per Paper	2.56	2.48	2.63	1.36	3.74
Average Number of Words	77537	29468	115791	84359	71006

Notes: Panel A restricts the sample to researchers in STEMM fields; Panel B to researchers in SSH fields. Within each panel, Column 1 reports the estimate for the full field-specific sample. Columns 2 and 3 split disciplines into Low- and High-Words groups at the median of the discipline-level mean total number of words written by evaluators for all candidates, calculated separately within STEMM and SSH. Columns 4 and 5 split disciplines by the median number of coauthors per paper. The estimates are based on Equation (3). The outcome is accumulated total research output during the first three years following the exam. All specifications control for the probability of being drawn, exam fixed effects, past research production, and a second-order polynomial in academic age. Committee membership is instrumented using the initial random draw. Standard errors are clustered at the exam level.
*** $p < .01$, ** $p < .05$, * $p < .1$.

Table 4: Impact of sitting on the committee on total research output by baseline research quality

	Accumulated Total Works 3 years post-exam		Accumulated Total Works 8 years post-exam	
	(1)	(2)	(3)	(4)
Drawn Researcher (Scaled)	-0.187***	-0.193***	-0.089	-0.092*
	(0.045)	(0.045)	(0.056)	(0.055)
Drawn Researcher (Scaled) x Evaluator Quality		-0.136**		-0.118
		(0.057)		(0.072)
Observations	15712	15712	9552	9552
Adjusted R^2	0.49	0.49	0.43	0.43
Number of Exams	630	630	347	347

Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. The estimates are based on Equation (4). In columns 1-2, the outcome variable is accumulated total research output in the first three years following the exam. In columns 3-4, the outcome variable is accumulated total research output in the first eight years following the exam. *Evaluator quality* (T_j) is the quality-adjusted publication record of evaluators over the ten years preceding the exam. We control for the probability of being drawn, exam fixed effects, past research production, and a second-order polynomial in academic age. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table 5: Probability of volunteering again

	Volunteer Again		
	(1)	(2)	(3)
Drawn Researcher (Scaled)	-0.008	-0.083**	-0.084***
	(0.032)	(0.032)	(0.032)
Evaluator Quality	0.019***	0.023***	0.020**
	(0.004)	(0.005)	(0.008)
Drawn Researcher (Scaled) × Evaluator Quality			-0.069**
			(0.034)
Observations	9552	8696	8696
Adjusted R^2	0.002	0.002	0.003
Restricted Sample		✓	✓
Exam Fixed Effects	✓	✓	✓

Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system in 2012 and 2016. In columns 2-3, the sample is restricted to potential evaluators in 2012 and 2016 who were not drawn for committee service in 2016 and 2018-2021, respectively. Excluding volunteers who were drawn in these intermediate cycles removes the mechanical upward bias that would otherwise arise from the draw bar, which mechanically reduces later participation in the counterfactual group. The outcome variable is an indicator for whether a researcher in the pool of potential evaluators volunteers again in the next cycle in which they are eligible to participate (the 2018 cycle for 2012 volunteers and the 2023 cycle for 2016 volunteers). *Evaluator quality* (T_j) is measured as the quality-adjusted publications of the evaluator over the past 10 years prior to the exam. We control for researchers' (re-centered) time since first publication, the (re-centered) probability of being drawn. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table 6: Committee composition and individual voting

	Positive Vote			
	(1)	(2)	(3)	(4)
Evaluator Quality	-0.023**	-0.000	-0.023**	-0.023**
	(0.011)	(0.002)	(0.011)	(0.011)
Peer Quality	-0.055**		-0.055**	-0.055**
	(0.024)		(0.024)	(0.024)
Evaluator Quality × Candidate's Number of Publications			-0.005**	-0.005
			(0.003)	(0.004)
Evaluators Quality × Candidate's Impact of Publications			0.006***	0.007**
			(0.002)	(0.004)
Peer Quality × Candidate's Number of Publications				-0.017*
				(0.010)
Peer Quality × Candidate's Impact of Publications				0.030***
				(0.009)
Observations	241744	241744	241744	241744
Adjusted R^2	0.13	0.00	0.13	0.13
Candidate-Evaluator Connections	✓	✓	✓	✓
Candidate Fixed Effects		✓		

Notes: The unit of observation is individual reports at the candidate-evaluator level. The outcome variables are textual features of the evaluation reports. All outcome variables are normalized at the macro field-category level. There are 86 macro fields and 2 categories: Associate and Full. *Evaluator quality* (T_j) is the quality-adjusted publication record of evaluator j over the ten years preceding the exam. *Peer quality* ($T_{e,-j}$) is the average quality-adjusted publication record of evaluators, excluding j, over the ten years preceding the exam. We control for the expected quality of peers, researchers' (re-centered) probability of being drawn, candidate productivity, and connections (coauthors, colleagues, same disciplinary sector) in all specifications. The quality of peers is instrumented by the initial quality of peers, due to potentially endogenous resignations. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

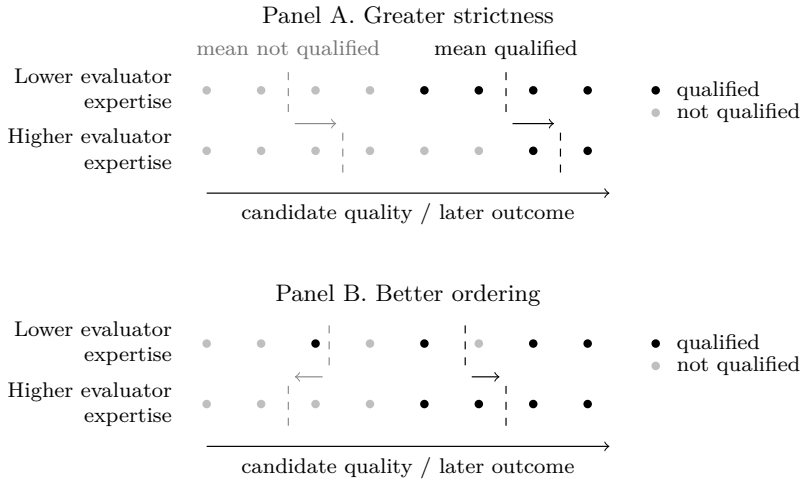
Table 7: Committee composition and individual report writing

	Words	Average IDF	Total IDF
	(1)	(2)	(3)
Evaluator Quality	-0.030	-0.002	-0.018
	(0.042)	(0.037)	(0.038)
Peer Quality	0.202**	0.169**	0.198**
	(0.099)	(0.084)	(0.097)
Observations	241744	241744	241744
R^2	0.020	0.005	0.017
Candidate-Evaluator Connections	✓	✓	✓
Candidate Controls	✓	✓	✓

Notes: The unit of observation is individual reports at the candidate-evaluator level. The outcome variables are textual features of the evaluation reports. All outcome variables are normalized at the macro field-category level. There are 86 macro fields and 2 categories: Associate and Full. *Evaluator quality* (T_j) is the quality-adjusted publication record of evaluator j over the ten years preceding the exam. *Peer quality* ($T_{e,-j}$) is the average quality-adjusted publication record of evaluators, excluding j, over the ten years preceding the exam. We control for the expected quality of peers, researchers' (re-centered) probability of being drawn, candidate productivity, and connections (coauthors, colleagues, same disciplinary sector) in all specifications. The quality of peers is instrumented by the initial quality of peers, due to potentially endogenous resignations. Standard errors are clustered at the exam level.
*** $p < .01$, ** $p < .05$, * $p < .1$

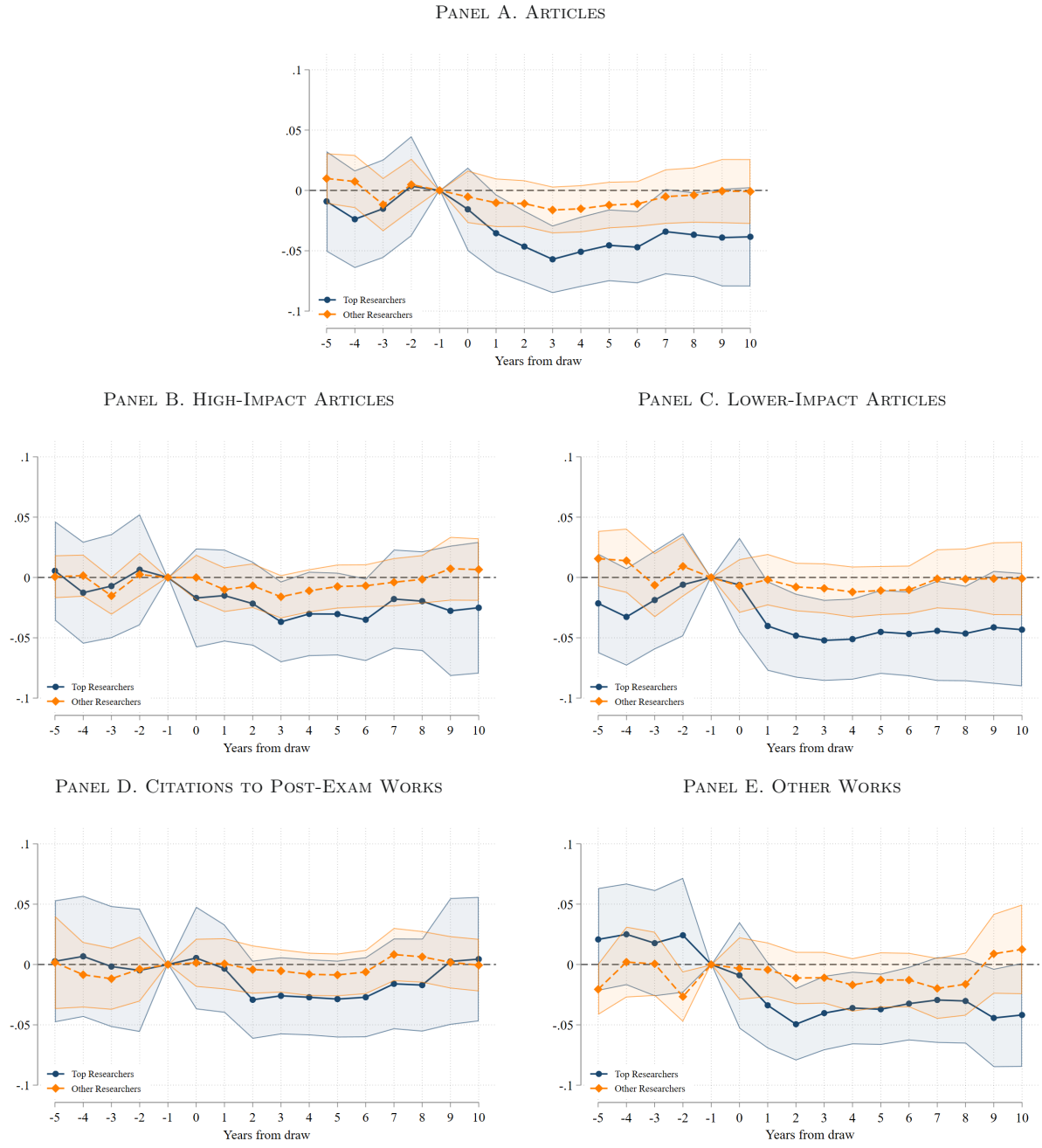
A Appendix Figures and Tables

Figure A1: Visual example of candidate outcomes under greater strictness and better ordering



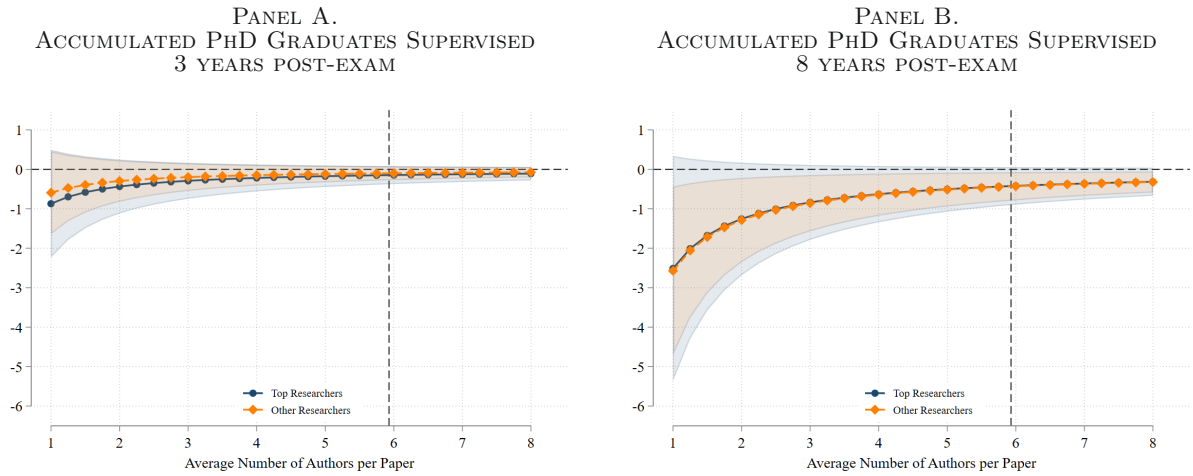
Notes: Each dot represents a candidate, ordered horizontally by candidate quality or by a later outcome that is increasing in candidate quality. Black dots denote qualified candidates and gray dots denote candidates who are not qualified. Panel A isolates greater strictness: fewer candidates are qualified, which raises the average outcome among qualified candidates but also raises the average outcome among non-qualified candidates because higher-quality marginal candidates are rejected. Panel B isolates better ordering at a fixed qualification rate: qualified candidates are drawn from higher ranks, raising their average outcome, while non-qualified candidates are drawn from lower ranks, lowering their average outcome.

FIGURE A2: IMPACT OF SITTING ON THE COMMITTEE ON ACCUMULATED ARTICLE PRODUCTION BY EVALUATORS' BASELINE PRODUCTIVITY



Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. The estimates are based on Equation (4). The outcome variable is accumulated articles starting from the year prior to the exam, normalized at the field-year level. The solid blue line shows estimates for top researchers. The orange dashed line shows estimates for other researchers. The displayed point estimates are computed for researchers with the average number of coauthors in the sample (5.93). *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*. *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. We control for the probability of being drawn, exam fixed effects, past research production, and researchers' time since first publication. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level. The shaded area denotes the 95% confidence interval around the estimates.

Figure A3: Impact of sitting on the committee on total number of PhD graduates supervised by top researcher over the distribution of collaborators per paper prior to the exam



Notes: The sample is all researchers who participate in the Italian evaluation system between 2012-2021. The estimates are based on Equation (4). The outcome variable is accumulated number of PhD students graduating under researchers' supervision starting from the year prior to the exam. We control for the probability of being drawn, and exam fixed effects. We instrument whether the researcher sits on the committee by the initial random draw. The solid blue line shows estimates for top researchers. The orange dashed line shows estimates for other researchers. The dashed line on the x-axis researchers' average number of authors per paper in the sample (5.93). *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. Standard errors are clustered at the exam level. The shaded areas denote the 95% confidence interval around the estimates.

Table A1: Evaluator characteristics by eligibility, volunteering, and resignation status

Panel A.
Eligibility

	Not Eligible		Eligible		Difference	
	Mean	SD	Mean	SD	Diff	P-value
Top Researcher	0.11	0.32	0.42	0.49	-0.30***	0.00
Total Works	-0.36	0.73	0.38	1.05	-0.73***	0.00
Articles	-0.36	0.73	0.37	1.05	-0.73***	0.00
High-Impact Publications	-0.37	0.69	0.37	1.08	-0.74***	0.00
Lower-Impact Publications	-0.25	0.79	0.28	1.08	-0.53***	0.00
Observations	23859		28544		52403	

Panel B.
Volunteering conditional on eligibility

	Not Volunteered		Volunteered		Difference	
	Mean	SD	Mean	SD	Diff	P-value
Top Researcher	0.38	0.48	0.47	0.50	-0.09***	0.00
Total Works	0.27	1.03	0.52	1.06	-0.25***	0.00
Articles	0.26	1.04	0.51	1.06	-0.25***	0.00
High-Impact Publications	0.28	1.06	0.48	1.10	-0.20***	0.00
Lower-Impact Publications	0.18	1.05	0.40	1.10	-0.22***	0.00
Observations	16100		12444		28544	

Panel C.
Resignation conditional on being drawn

	Not Resigned		Resigned		Difference	
	Mean	SD	Mean	SD	Diff	P-value
Top Researcher	0.40	0.49	0.45	0.50	-0.05*	0.08
Total Works	0.30	1.02	0.41	1.04	-0.11*	0.07
Articles	0.30	1.01	0.40	1.04	-0.10*	0.10
High-Impact Publications	0.28	1.05	0.41	1.17	-0.13*	0.06
Lower-Impact Publications	0.25	1.04	0.30	0.99	-0.05	0.39
Observations	3085		331		3416	

Notes: Panel A. contains the sample of Italian Full Professors at public universities in the years prior to the exam in exam years: 2012, 2016, 2018, 2021. Panel B. contains the sample of Italian Full Professors who meet eligibility criteria to participate as evaluators. Panel C. contains the sample of Italian Full Professors who were initially drawn to sit on a committee. The productivity measures are accumulated for 10 years up to the year prior to the exam, and normalized at the exam 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*. *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. Because eligibility is not directly observed, we simulate it following the official criteria published by the Ministry.
*** $p < .01$, ** $p < .05$, * $p < .1$

Table A2: Research productivity in the pool of potential evaluators prior to the exam

	Overall		Top Researchers		Other Researchers	
	Mean	SD	Mean	SD	Mean	SD
Works	60	70	84	91	45	48
Articles	53	63	75	82	40	42
High-Impact Articles	22	34	37	46	13	18
Lower-Impact Articles	31	36	38	44	27	30
PhD Students Supervised	3	6	4	7	3	6
Observations	16683		6402		10281	
Number of Exams	685		685		685	
Number of Researchers	10487		4344		4344	

Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. Panel B. and Panel C. splits the sample between top and other researchers. All research output measures are counted in the ten years prior to the exam. The number of PhD students supervised is only defined for a subset of researchers, whose institutions consistently reported their graduates. *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*. *Top Researcher* is defined as 75th percentile among all Italian Full Professors based on quality-adjusted publications in the last 10 years prior to the exam.

Table A3: Research productivity in the pool of potential evaluators following the exam

	Overall		Top Researchers		Other Researchers	
	Mean	SD	Mean	SD	Mean	SD
Year 0						
Works	7	11	10	14	5	7
PhD Students Supervised	0	1	1	2	0	1
Years 0–3 (cumulative)						
Works	28	39	38	52	21	27
Articles	24	36	34	48	19	25
High-Impact Articles	9	18	15	25	6	11
Lower-Impact Articles	15	21	19	27	13	17
PhD Students Supervised	2	5	2	6	2	4
Years 0–8 (cumulative)						
Works	50	79	71	102	36	55
Articles	45	74	63	95	32	51
High-Impact Articles	18	37	28	48	11	23
Lower-Impact Articles	27	42	35	52	21	32
PhD Students Supervised	4	8	4	9	3	7
Observations	16683		6402		10281	

Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. Panel B. and Panel C. splits the sample between top and other researchers. The number of PhD students supervised is only defined for a subset of researchers, whose institutions consistently reported their graduates. *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*. *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.

Table A4: The impact of committee quality on application withdrawal

	Application Withdrawn
	(1)
Candidate's Number of Publications	-0.039***
	(0.002)
Candidate's Impact of Publications	-0.027***
	(0.003)
Committee Quality	0.012
	(0.013)
Candidate's Number of Publications × Committee Quality	-0.003
	(0.005)
Candidate's Impact of Publications × Committee Quality	-0.004
	(0.008)
Observations	69020
Number of Exams	184
R^2	0.03
Candidate Connections	✓

Notes: The sample is all candidates who submitted an application in the first wave of the 2012 cycle of the Italian evaluation system. The outcome variable is a dummy indicating whether the candidate withdrew their application following the public announcement of the committee and evaluation criteria. *Committee quality* (T_e) s the average quality-adjusted publication record of evaluators over the ten years preceding the exam. The final composition of the committee is instrumented by the initial committee composition, before resignations took place. All candidate productivity measures are normalized within each exam, and evaluator quality is normalized within each eligibility pool. The final composition of the committee is instrumented by the initial committee composition, due to potentially endogenous resignations. We control for expected committee quality, its interactions with candidates' publication quantity and impact, and the number of connections between evaluators and candidates (coauthors, colleagues, same disciplinary sector). Standard errors are clustered at the field level.
*** p < .01, ** p < .05, * p < .1

Table A5: Candidates in the 2012 exams

	Overall		By Field				By Application Level			
			STEM		SSH		ASP		FP	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Year of First Publication	1997	8	1996	8	1998	8	1998	7	1992	8
Works	46	42	56	47	31	27	41	37	58	51
Articles	25	31	35	35	11	14	22	26	33	38
High-impact Articles	10	17	16	20	3	5	9	14	14	21
Observations	69020		41395		27625		47426		21594	
Number of Candidates	46329		27613		19388		34643		16747	

Notes: The sample is all candidates who participated in the first wave of the 2012 cycle of evaluations. The productivity measures are constructed from candidates' curricula vitae. *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*.

Table A6: Future productivity of successful candidates in the 2012 exams

	Panel A. Overall				Panel B. Economics Fields			
	All Qualified		Top-ranked		All Qualified		Top-ranked	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total Works	62	92	63	93	32	33	32	33
Articles	56	86	57	88	29	31	29	30
High-Impact Articles	24	45	24	46	10	11	10	11
Citations: Pre-Exam Works	1029	2192	1050	2231	385	644	400	653
Citations: Post-Exam Works	1475	5783	1514	5863	510	1399	520	1431
Observations	25342		23776		2236		2092	

Notes: The sample is candidates who qualified in the first wave of the 2012 cycle of evaluations. The productivity measures are accumulated for 10 years from the year of the exam. Pre-exam citations are measured between 2012-2022 for works published up to 2012. Post-exam citations are measured between 2012-2022 for works published in 2012 or later. 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 and C displays the sample of all qualified candidates. In Panel B and D we restrict the sample to include candidates qualified unanimously by committees where less-productive researchers are drawn than expected – this equalizes the share of successful candidates across exams.

Table A7: Evaluation reports

	Overall		STEM		SSH		Difference	
	Mean	SD	Mean	SD	Mean	SD	Diff	P-value
Positive Vote	0.45	0.50	0.46	0.50	0.43	0.49	0.03	0.000
Words	182.45	285.88	150.32	192.24	232.78	384.32	-82.46	0.000
Average IDF	2.54	0.76	2.29	0.54	2.94	0.87	-0.65	0.000
Total IDF	304.87	619.88	219.98	389.98	437.82	847.89	-217.84	0.000
Any Bibliometric Mention	0.45	0.50	0.47	0.50	0.43	0.50	0.04	0.000
Observations	241744		147538		94206		241744	
Number of Evaluators	758		450		308			
Number of Candidates	40217		23953		16769			

Notes: The sample is all evaluation reports in the first wave of the 2012 cycle of evaluations. *Words* measures the number of words included in the evaluation report. *Average IDF* is the average *inverse-document-frequency* score of the words in the evaluation report. *Total IDF* is *Words* $\times$ *Average IDF*. Bibliometric indicators include *h-index, citations, A-list articles, Scopus, Web of Science, etc.*

Table A8: Report characteristics and candidate profiles

	Words			Average IDF			Total IDF		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coauthors	0.123**	0.094***	0.115***	0.135***	0.106***	0.117***	0.148***	0.114***	0.134***
	(0.057)	(0.028)	(0.020)	(0.039)	(0.031)	(0.027)	(0.053)	(0.028)	(0.021)
Same University	0.109***	0.090**	0.089***	0.050	0.068**	0.072***	0.106***	0.094**	0.100***
	(0.039)	(0.037)	(0.019)	(0.035)	(0.034)	(0.024)	(0.038)	(0.037)	(0.019)
Same Subfield	0.169***	0.112***	0.128***	0.050*	0.115***	0.093***	0.148***	0.120***	0.138***
	(0.032)	(0.039)	(0.025)	(0.029)	(0.037)	(0.023)	(0.031)	(0.039)	(0.028)
Candidate's Number of Publications	0.057***			0.017*			0.058***		
	(0.009)			(0.009)			(0.009)		
Candidate's Impact of Publications	0.038***			0.029***			0.043***		
	(0.010)			(0.007)			(0.009)		
CVs to Evaluate (100s)	-0.008	-0.011		-0.023**	-0.051***		-0.014	-0.026	
	(0.019)	(0.027)		(0.012)	(0.018)		(0.018)	(0.025)	
Observations	241744	241744	241744	241744	241744	241744	241744	241744	241744
Adjusted R^2	0.013	-0.197	-0.192	0.006	-0.190	-0.196	0.013	-0.195	-0.192
Candidate Fixed Effect		✓	✓		✓	✓		✓	✓
Evaluator Fixed Effect			✓			✓			✓

Notes: The sample is all evaluation reports in the first wave of the 2012 cycle of evaluations. The outcome variables are textual features of the evaluation reports. All measures are normalized at the macro field-category level. There are 86 macro fields and 2 categories: Associate and Full. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A9: Report characteristics and candidate outcomes

	Total Words			Average IDF			Total IDF		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Marginal Candidate	0.203***	0.167***	0.169***	0.007	-0.014	-0.026	0.191***	0.153***	0.161***
	(0.057)	(0.057)	(0.039)	(0.042)	(0.043)	(0.035)	(0.053)	(0.053)	(0.036)
Observations	241744	241744	241744	241744	241744	241744	241744	241744	241744
Adjusted R^2	0.009	0.015	0.603	0.000	0.002	0.450	0.008	0.014	0.553
Candidate Controls		✓	✓		✓	✓		✓	✓
Evaluator Fixed Effect			✓			✓			✓

Notes: The sample is all evaluation reports in the first wave of the 2012 cycle of evaluations. The outcome variables are textual features of the evaluation reports. All measures are normalized at the macro field-category level. There are 86 macro fields and 2 categories: Associate and Full. *Marginal Candidate* is a dummy variable indicating whether a candidate got 3 or 4 positive votes. *Candidate Controls* includes the number of publications and the share of high-impact articles. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A10: Randomization check on candidate characteristics

	(1) Publcations	(2) Share High-Impact	(3) Articles	(4) High-Impact	(5) AIS
Committee Quality	-4.795	-0.013	-3.374	-2.265	-5.919
	(3.705)	(0.025)	(3.443)	(1.905)	(6.138)
Expected Committee Quality	22.700	0.073	0.514	-4.541	-19.084
	(30.167)	(0.320)	(28.281)	(20.176)	(58.315)
Observations	69020	67323	69020	69020	69020

Notes: The dependent variables are measures of candidate research productivity prior to the exam. *High impact articles* are 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.

Table A11: First stage results from committee-level regression

	Final Committee Quality (1)
Random Shock to Initial Committee Quality ($\tilde{Z}_e$)	0.834***
	(0.074)
Candidate's Number of Publication	0.000
	(0.000)
Candidate's Impact of Publications	-0.000
	(0.000)
Observations	69020
KP rk Wald F-stat	128
Number of Exams	184

Notes: The table reports the first stage regression from Equation (1). The first stage corresponds to the model in column 1 of Table 1. *Initial Committee Quality* is the average quality-adjusted publication record of evaluators who were initially drawn over the ten years preceding the exam. Evaluator quality is standardized within each eligibility pool. We control for the expected committee quality, and broad discipline (4 categories) and application level (Full and Associate). Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

TABLE A12: EVALUATOR QUALITY AND THE PROBABILITY OF MENTIONING BIBLIOMETRIC INDICATORS IN CANDIDATE REPORTS

	Mentioning Bibliometric Indicator			
	(1)	(2)	(3)	(4)
	Overall	Overall	Negative Vote	Positive Vote
Evaluator Quality	-0.028**	-0.025*	-0.032**	-0.016
	(0.014)	(0.013)	(0.013)	(0.014)
Observations	241669	241669	133450	108219
Adjusted R^2	0.010	0.075	0.057	0.094
Candidate Connections	✓	✓	✓	✓
Report Length		✓	✓	✓

Notes: The sample is all evaluation reports in the first wave of the 2012 cycle of evaluations. The outcome variable is a binary indicator equal to one if bibliometric indicators are mentioned in the evaluation report. Bibliometric indicators include *h-index, citations, A-list articles, Scopus, Web of Science, etc. Evaluator quality* (T_j) is the quality-adjusted publication record of the evaluator over the ten years preceding the exam. Evaluator quality is standardized within each eligibility pool. We control for the expected quality of the committee. Standard errors are clustered at the exam level. *** p < .01, ** p < .05, * p < .1

Table A13: Committee composition and the predictive content of qualification decisions (without controls)

	All Candidates					Associate Professor Candidates		
	(1) Total Publications	(2) High-Impact Articles	(3) Citations (All Articles)	(4) Citations (Pre-Exam Articles)	(5) Citations (Post-Exam Articles)	(6) Promoted Full Prof.	(7) Future Evaluator	(8) Future Top Evaluator
Panel A. Successful Candidates								
Committee Quality	0.023	0.058**	0.051*	0.044	0.042*	0.054***	0.040***	0.015***
	(0.030)	(0.028)	(0.029)	(0.030)	(0.024)	(0.016)	(0.010)	(0.005)
Observations	25342	25342	25342	25342	25342	17541	17541	17541
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	0.25	0.27	0.15	0.19	0.11	0.17	0.10	0.04
Panel B. Unsuccessful Candidates								
Committee Quality	0.026	0.007	-0.008	0.002	-0.009	0.010*	0.004	0.001
	(0.017)	(0.021)	(0.019)	(0.021)	(0.015)	(0.005)	(0.004)	(0.002)
Observations	43678	43678	43678	43678	43678	29885	29885	29885
R^2	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	-0.14	-0.16	-0.09	-0.11	-0.06	0.04	0.02	0.01

Notes: The table reports estimates of Equation (2). The dependent variables measure candidate outcomes ten years after the evaluation. In Panel A. the sample is all candidates who submitted an application in the first wave of the 2012 cycle of the Italian evaluation system. In Panel B. we restrict the sample to candidates to Associate Professor applicants in the first wave of the 2012 cycle. *Promoted Full* is a dummy variable indicating whether the candidate is promoted to Full Professor within ten years. *Qualified* is an indicator equal to one if the candidate is qualified. *Committee Quality* is defined as the deviation of realized committee quality from its expected value under the assignment mechanism, measured as the average quality-adjusted publication record of evaluators over the ten years preceding the exam. All candidate productivity measures are normalized within each exam, and evaluator quality is normalized within each eligibility pool. Both committee quality and candidate outcome variables are normalized at the examination level. All specifications include exam fixed effects and candidate controls for affiliation and academic rank. Standard errors are clustered at the examination level.
*** p < .01, ** p < .05, * p < .1

Table A14: Effect of Committee Quality on Post-Exam Outcomes among Qualified Candidates, Unanimity-Restricted Sample

	All Candidates					Associate Professor Candidates		
	(1) Total Publications	(2) High-Impact Articles	(3) Citations (All Articles)	(4) Citations (Pre-Exam Articles)	(5) Citations (Post-Exam Articles)	(6) Promoted Full Prof.	(7) Future Evaluator	(8) Future Top Evaluator
Panel A. Unanimity Restricted Successful Candidates (without Controls)								
Share of candidates qualified by better than average committee: 0.341								
Share of candidates unanimously qualified by worse than average committee: 0.338								
Above Average Committee	-0.001	0.027	0.046**	0.044*	0.032	0.035**	0.023**	0.006
	(0.024)	(0.025)	(0.023)	(0.025)	(0.020)	(0.014)	(0.009)	(0.005)
Observations	23424	23424	23424	23424	23424	16269	16269	16269
R^2	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	0.26	0.28	0.16	0.20	0.12	0.17	0.11	0.04
Panel B. Unanimity Restricted Successful Candidates (with Controls)								
Share of candidates qualified by better than average committee: 0.341								
Share of candidates unanimously qualified by worse than average committee: 0.338								
Above Average Committee	-0.004	0.009	0.025*	0.010	0.023	0.027**	0.020**	0.005
	(0.017)	(0.019)	(0.015)	(0.012)	(0.015)	(0.012)	(0.008)	(0.004)
Observations	23420	23420	23420	23420	23420	16266	16266	16266
R^2	0.18	0.16	0.37	0.61	0.16	0.02	0.01	0.02
Baseline	0.26	0.28	0.16	0.20	0.12	0.17	0.11	0.04

Notes: The table reports OLS estimates of the effect of being initially assigned a committee of an above-expected quality on post-exam outcomes among top-qualified candidates. Top-qualified candidates are those qualified in committees with a positive shock to committee quality and those qualified by unanimity in committees with a negative shock to committee quality. Panel A includes discipline fixed effects. Panel B additionally includes candidate productivity controls: publication, share of high-impact articles and citations as observed at the moment of the exam call. Standard errors are clustered at the exam level. Columns (1)–(5) are estimated on all top-qualified candidates. Column (6) is estimated on Associate Professor candidates only.
*** p < .01, ** p < .05, * p < .1

Table A15: Randomization check on evaluator characteristics

	(1) Publications	(2) Articles	(3) High-impact Articles	(4) Citations	(5) Academic age
Drawn researcher	-0.009	-0.012	0.001	0.013	0.207
	(0.024)	(0.024)	(0.024)	(0.025)	(0.207)
Observations	15712	15712	15712	15712	15712

Notes: The dependent variables are evaluator observable characteristics prior to the exam. Productivity measures are accumulated ten years prior to the exam and normalized at the field level. *High impact articles* are articles in top-quartile Web of Science journals (STEM+M) or A-list journals (SSH). We instrument the final draw status with the random component of the initial draw. Standard errors are clustered at the exam level.

Table A16: First stage results from evaluator-level regression

	Final Draw Status
	(1)
Random Shock to Initial Draw Status ($\tilde{Z}_{j,e}$)	0.888***
	(0.007)
Number of Publications	0.001
	(0.002)
High-Impact Articles	-0.003
	(0.002)
Observations	15712
KP rk Wald F-stat	16497
Number of Exams	685

Notes: The table reports the first stage regression from Equation (3). The first stage corresponds to the model in column 1 of Figure 2. *Initial Committee Quality* is the average quality-adjusted publication record of evaluators who were initially drawn over the ten years preceding the exam. Evaluator quality is standardized within each eligibility pool. We control for the expected committee quality, and broad discipline (4 categories) and application level (Full and Associate). Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A17: Horse-race regression between indicator and scaled draw status

	Accumulated Total Works 3 years post-exam		
	(1)	(2)	(3)
Drawn Researcher	-0.051***		0.007
	(0.017)		(0.026)
Drawn Researcher (Scaled)		-0.187***	-0.202***
		(0.045)	(0.070)
Observations	15712	15712	15712
Adjusted R^2	0.49	0.49	0.49

Notes: The outcome variable is accumulated total research output in the first three years following the exam. The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system between 2012-2021. We control for the probability of being drawn, exam fixed effects, past research production, and a second-order polynomial in academic age. We instrument whether the researcher sits on the committee by the initial random draw. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A18: Non-linear effect of top researchers on evaluation outcomes

	Qualified	
	(1)	(2)
At Least One Top Researcher on Committee	-0.058*	
	(0.032)	
One Top Researchers on Committee		-0.042
		(0.039)
Two Top Researchers on Committee		-0.073**
		(0.036)
Three or More Top Researchers on Committee		-0.086**
		(0.039)
Observations	69020	69020
Adjusted R^2	0.092	0.096

Notes: The sample is all candidates who submitted an application in the first wave of the 2012 cycle of the Italian evaluation system. In columns 1 and 2, the dependent variable is a binary indicator equal to one if the candidate qualifies. *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. We control for the probability to draw one, two, and three or more top researchers, respectively. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A19: Non-linear Impact of Top Peers on Evaluator Report Writing

	Words	Average IDF	Total IDF
	(1)	(2)	(3)
One Top Researcher Peer	0.298**	0.089	0.289**
	(0.138)	(0.139)	(0.135)
Two Top Researcher Peers	0.412***	0.112	0.362***
	(0.121)	(0.137)	(0.118)
Three or More Top Peers	0.291**	0.184	0.292**
	(0.131)	(0.140)	(0.126)
Observations	241744	241744	241744
Adjusted R^2	0.028	0.006	0.019
Candidate-Evaluator Connections	✓	✓	✓
Candidate Controls	✓	✓	✓

Notes: The sample is all evaluation reports in the first wave of the 2012 cycle of evaluations. The outcome variables are textual features of the evaluation reports. All measures are normalized at the macro field-category level. There are 86 macro fields and 2 categories: Associate and Full. *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. We control for the probability to draw one, two, and three or more top peer evaluators, respectively. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table A20: Non-linear effect of top researchers on future outcomes of qualified candidates

	Total Publications		Share High-Impact		Citations (Pre-exam works)		Citations (Post-exam works)		AP Candidate Promoted FP	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
One Top Researchers on Committee	0.044	0.018	0.016	0.003	0.027	-0.021	0.054*	0.019	0.012	0.024
	(0.045)	(0.031)	(0.029)	(0.022)	(0.027)	(0.015)	(0.029)	(0.026)	(0.022)	(0.022)
Two Top Researchers on Committee	0.022	-0.002	0.031	0.008	0.037	-0.007	0.051*	0.018	0.026	0.033
	(0.045)	(0.031)	(0.030)	(0.022)	(0.028)	(0.017)	(0.028)	(0.025)	(0.024)	(0.023)
Three Top Researchers on Committee	0.049	0.014	0.042	0.004	0.081**	0.009	0.069**	0.026	0.063**	0.064**
	(0.048)	(0.033)	(0.031)	(0.023)	(0.031)	(0.016)	(0.031)	(0.025)	(0.024)	(0.025)
Four Top Researchers on Committee	0.066	0.020	0.157***	0.101***	0.093***	0.041*	0.082*	0.057	0.077**	0.054**
	(0.064)	(0.045)	(0.039)	(0.026)	(0.027)	(0.022)	(0.047)	(0.046)	(0.035)	(0.027)
Obs.	25342	25337	25342	25337	25342	25337	25342	25337	17541	17536
Adj. R^2	0.004	0.195	0.002	0.195	0.009	0.601	0.005	0.157	0.011	0.118
Pre-Exam Observables		✓		✓		✓		✓		✓

Notes: The sample is qualified candidates in exams where realized committee quality is above its expectation, but restricted to unanimously qualified candidates in exams where realized committee quality is below expectation. The outcome variables are accumulated for 10 years from the year prior to the exam, and normalized at the exam level. Pre-exam citations (columns 5 and 6) are measured for works published before 2012. Post-exam citations (columns 7 and 8) are measured for works published in 2012 or later. 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. We control for the probability to draw one, two, and three or more top researchers, respectively. The *pre-exam observables* include the candidates' affiliation, academic rank, number of publications, share of high-impact articles, and total citations prior to the exam. Standard errors are clustered at the exam level.

*** p < .01, ** p < .05, * p < .1

Table A21: Productivity cost of sitting on committees by quality of fellow evaluators

	Panel A. Accumulated Total Works 3 years post-exam				Panel B. Volunteer Again			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Drawn Researcher (Scaled)	-0.187***	-0.172***	-0.175***	-0.175***	-0.077**	-0.086**	-0.067	-0.067
	(0.045)	(0.057)	(0.057)	(0.057)	(0.032)	(0.043)	(0.045)	(0.045)
Evaluator Quality	0.058***	0.103***	0.102***	0.103***	0.004	-0.001	-0.004	-0.004
	(0.013)	(0.015)	(0.015)	(0.015)	(0.008)	(0.011)	(0.011)	(0.011)
Drawn Researcher (Scaled) × Evaluator Quality		-0.135**	-0.132**	-0.129**		-0.074**	-0.073**	-0.073**
		(0.057)	(0.055)	(0.054)		(0.035)	(0.035)	(0.035)
Drawn Researcher (Scaled) × Peer Quality				0.035				0.012
				(0.096)				(0.076)
Observations	15712	15712	15712	15712	8696	8696	8696	8696
Adjusted R^2	0.489	0.490	0.490	0.490	0.031	0.031	0.037	0.037
Exam Fixed Effects	✓	✓			✓	✓		

Notes: The sample is all researchers who were in the pool of potential evaluators of the Italian evaluation system in 2012 and 2016. In Panel B. the sample is restricted to potential evaluators in 2012 and 2016 who were not drawn for committee service in 2016 and 2018-2021, respectively. Excluding volunteers who were drawn in these intermediate cycles removes the mechanical upward bias that would otherwise arise from the draw bar, which mechanically reduces later participation in the counterfactual group. In Panel A. the outcome measure is a sum of accumulated research output over the 3 years following the year of the exam, normalized at the exam-level. In Panel B. the outcome measure is a dummy variable, indicating whether the researcher volunteered again in subsequent cycles. Evaluator quality (T_j) is measured as the quality-adjusted publications of the evaluator over the past 10 years prior to the exam. The quality of the peers ($T_{e,-j}$) is measured as the average quality-adjusted publications of fellow evaluators over the past 10 years prior to the exam. We control for broad discipline (4 categories), the probability of being drawn and its interactions with expected peer quality. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

B Data construction

We construct the main database by combining four sources: (i) CINECA records on professors employed at Italian public universities, (ii) official ASN records on eligible evaluators, and committee draws, (iii) publication records from IRIS and OpenAlex, and (iv) candidate-level data collected by Bagues et al. (2017). This appendix summarizes the construction of researcher identifiers, the linkage of publication records, and the simulation of the ASN draw procedure.

Researcher identifiers. CINECA provides the backbone of the researcher database because it covers the universe of professors employed at Italian public universities. The data include each professor's name, gender, university, department, research area, and academic rank. We assign each professor a unique `cineca_id` that tracks individuals over time. In almost all cases, names uniquely identify professors within a year. For the small number of homonyms, we disambiguate using research area and, where necessary, university affiliation; the remaining cases are checked manually.

We then link ASN evaluator and candidate records to CINECA. Evaluators are matched primarily by name within macro field, after standardizing names and correcting common typographical and OCR inconsistencies in the Ministry files. Because eligible evaluators are drawn from the population of Italian Full Professors within each field, the risk of false matches is limited. This procedure links virtually all eligible evaluators to CINECA; only eight researchers in the evaluator pool remain unmatched.

Candidates are matched to CINECA using a more conservative procedure, because some candidates apply from outside the Italian university system and are therefore not observed in CINECA. We standardize names as for evaluators, but exclude riskier match types and require a match on both given and family names. We also exclude Full Professors from the set of potential CINECA matches, since ASN candidates apply for associate or Full Professorship qualification. This procedure links 33,446 of 46,329 candidates, or approximately 72 percent, to unique CINECA identifiers. Unmatched candidates are mainly researchers outside the Italian public university system.

Finally, we construct a master researcher identifier, `id`, that covers individuals appearing in CINECA, the ASN evaluator files, or the ASN candidate files. When a researcher is observed in multiple sources, we prioritize the CINECA identifier, then the evaluator identifier, and then the candidate identifier. We harmonize name, gender, affiliation, and field information within each `id`,

retaining one standardized researcher record for merging with publication data.

Publication records. We link researchers to publication data from IRIS and OpenAlex. IRIS is maintained by Italian universities and contains self-reported publication records of researchers affiliated with Italian institutions. We use two scrapes, from 2020 and 2024, and remove duplicate author-publication records within and across scrapes. IRIS author profiles are matched to researchers using standardized names and university affiliations. Exact and unique matches are accepted directly; non-unique names are resolved using affiliation histories. Ambiguous matches are discarded.

We complement IRIS with OpenAlex, which provides broader publication coverage and richer bibliographic metadata. The OpenAlex linkage begins with a broad set of name-based candidate matches, allowing for common differences in name order, initials, middle names, particles, and compound surnames. We then retain only links supported by additional evidence. Candidate pairs are confirmed when the name match is unique and is supported by at least one independent signal, such as affiliation history, journal overlap, country, or discipline. Confirmed OpenAlex profiles are removed from the candidate set for all other researchers before weaker rules are applied, preventing the same profile from being assigned to multiple individuals. Ambiguous or implausible links are discarded.

We combine IRIS and OpenAlex into a unified publication database. We harmonize publication types and retain substantive outputs, including journal articles, books, book chapters, and conference proceedings. Duplicate publications across sources are identified using cleaned titles, publication years, DOIs, ISSNs, journal names, and overlapping author information. When the same publication appears in both databases, we retain the OpenAlex record because it typically contains richer metadata and author identifiers. The resulting database contains one record per publication and forms the basis for the productivity and impact measures used in the analysis.

Simulation of ASN draw probabilities. We simulate the ASN draw procedure to recover each eligible evaluator's *ex ante* probability of being selected and the *ex ante* probability that any pair of evaluators serves on the same committee. These probabilities are used to control for expected committee composition and expected peer quality.

For the 2012 ASN, the draw procedure imposed two main constraints: representation across research areas and at most one evaluator per university. SSDs were ordered by the number of

eligible volunteers, with smaller SSDs considered first. One evaluator was drawn from each large SSD, defined as an SSD with more than 30 researchers in CINECA, subject to the university constraint. Remaining committee seats were then filled from the full pool of eligible evaluators, again respecting the constraint that no two selected evaluators came from the same university.

For the 2016–2023 ASN rounds, the same principles applied, but the rules for SSD representation changed. The threshold for guaranteed representation was reduced to 10 eligible Full Professors, and the number of guaranteed seats became proportional to SSD size. For each discipline, we compute the expected number of evaluators from SSD s as

$$E_s = \frac{N_s}{\sum_{s' \in S} N_{s'}} \times N,$$

where N_s is the number of eligible professors in SSD s, S is the set of SSDs in the discipline, and N is the number of evaluators drawn for the exam. The algorithm first fills the guaranteed SSD-specific seats, respecting the university constraint, and then fills any remaining seats from the full pool of eligible evaluators.

For each ASN round, we replicate the relevant draw algorithm one million times. The simulated draw probabilities are used in the empirical analysis to separate realized random variation in committee composition from predictable differences generated by the structure of the eligible pool and the draw rules.

C Robustness Checks: Alternative Measures of Research Profiles

In this section, we validate the publication database compiled from OpenAlex and IRIS by comparing it to candidates' self-reported curricula vitae, 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.

C.1 Validating OpenAlex and IRIS database

We begin by reporting summary statistics on candidate research profiles. We find about 92% of candidates in OpenAlex and IRIS. Candidates listed a larger number of works on their curricula vitae overall. This appears to reflect the fact that OpenAlex and IRIS provide less comprehensive coverage of non-journal research output. Specifically, we fail to capture the number of books and chapters written by candidates.

Next, we report the correlation between the measures coming from curricula vitae and the OpenAlex+IRIS database in Table C1. We find high correlations for journal-based output measures. For overall article production we report a correlation of 80%. Similarly, we see a correlation of 81% for high-impact articles. We find a positive but low correlation for books (5%) and chapters (6%). This is consistent with OpenAlex and IRIS being worse at tracking non-journal based forms of scientific output.

C.2 Candidate research profiles

In the main analysis, we characterize candidates' research output along two core dimensions: *quantity* and *impact*. 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. They are also only moderately correlated, ensuring that they provide complementary information. Quantity reflects the scale of a candidate's scholarly output, while impact aims to capture the perceived scientific contribution and value of this output.

The restriction to these two dimensions is motivated by both conceptual clarity and empirical considerations. First, impact 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, thematic breadth, or collaboration patterns – are often highly correlated with these two dimensions. Including them in the main analysis would therefore add complexity without

substantively altering the interpretation of results.

To situate our main measures within the broader landscape of research characteristics, we extend the analysis below to alternative indicators of impact and to additional attributes of researchers' profiles. The measures examined are listed below.

Measures used in the main analysis

Quantity. Total number of publications authored by the candidate prior to the exam. To account for systematic differences in scholarly communication across fields, we include journal articles, books, book chapters, and conference proceedings.

Impact (Journal-based). Share of a candidate's articles published either in the top quartile of Web of Science journal rankings (STEM+M fields) or in A-list journals designated by the ministry (Social Sciences and Humanities). This reflects the perceived selectivity and impact of publication outlets.

Alternative measures of impact

Impact (Article Influence Score). An alternative journal-impact measure based on the *Article Influence Score* (AIS), which captures the average influence of articles in a journal over the five years following publication. AIS is derived from the Eigenfactor metric and adjusts for field differences in citation practices, self-citations, and the impact of citing sources.

Impact (Citation-based). Average number of citations per publication, reflecting the scholarly recognition and influence of a candidate's work.

Additional profile characteristics

Measures related to perceived contribution to coauthored work:

Collaborators per Paper. Average number of coauthors per publication, indicating the breadth of a candidate's collaboration network.

First+Last+Single Authored Share. Share of publications where the researcher appears as first author, last author, or sole author—positions typically associated with leadership or primary intellectual contribution.

Measures related to career stage:

Seniority. Number of years since the candidate's first recorded publication, capturing accumulated research experience.

Measures related to thematic content:

Topic Dispersion. Number of distinct Open Alex topics in which the candidate has published, divided by total publications. This measures thematic concentration versus diversification.

Novelty. Index capturing the extent to which a candidate's research introduces new combinations of words or phrases in the literature (Arts et al., 2025). Following evidence of its unreliability for Social Sciences and Humanities, we restrict this measure to STEM+M fields.

Tables C3 and C4 summarize the correlations among these dimensions and their association with qualification outcomes. All measures are standardizes for candidates within the same exam. Individually, all measures correlate with qualification in the expected direction: quantity, journal impact, citations, seniority, and novelty show positive associations, whereas the number of collaborators and topic dispersion correlate negatively. The share of first-, last-, or single-authored publications shows no unconditional relationship with success, likely due to lower chances of occupying these positions in large, high-impact collaborations.

Among all indicators, *Quantity* and *Journal Impact* display the strongest predictive power (as measured by the adjusted R-squared), supporting their use as the core dimensions in the main analysis. In conditional specifications – including novelty for STEM+M (column 10) and excluding it for all fields (column 11) – the journal-based impact measure remains dominant: its coefficient increases, while the effect of citations attenuates substantially, and AIS becomes indistinguishable from zero in the full sample (and small and negative in STEM+M fields). Topic diversity appears to be another important predictor of success, with higher diversity negatively correlated with lower qualification rates.

C.3 Committee quality and returns to different dimensions of candidate research profiles

Following the logic of the analysis in Section 4, we examine whether committee quality affects the estimated returns to additional dimensions of candidates' research profiles, beyond the quantity and impact measures analyzed in the main text. Table C5 reports the results. Consistent with the main analysis, committee quality appears to significantly amplify the returns to publication impact. In contrast, the returns to publication quantity are no longer significantly affected by committee quality, likely reflecting a correlated (marginally significant) negative effect of committee quality on seniority, which is now controlled for in the regression. The returns to other dimensions of candidates' research profiles, conditional on these effects, do not appear to be significantly influenced by committee quality.

C.4 Robustness to the choice of measures of candidate publication quantity and impact

We begin by confirming that our results (reproduced in Column 1 of Table C6) are robust to using an alternative data source. Column 2 of Table C6 reports the results of the same analysis based on OpenAlex and IRIS publication data, rather than curricula vitae. Using this dataset explains less variation in the outcome variable, as reflected in a lower R^2, and yields a slightly smaller estimated return to publication impact. This is likely because the alternative data no longer exploit information directly provided to evaluators. The key coefficient of interest – the interaction between *Candidate's Publication Impact* and *Committee Quality* – is statistically indistinguishable from the estimate in the main text, with point estimates differing by only 0.2 percentage points.

Column 3 reports results using an alternative measure of *Quantity*, scaled by the candidate's average number of coauthors. This adjustment accounts for the possibility that evaluators place less weight on publications produced by larger teams. The R^2 is lower in this specification, suggesting that this variable is less informative about evaluators' decision-making process.

Next, we test the robustness of our findings to alternative measures of publication impact. First, we assess whether defining impact as the *share* of high-impact articles drives our results. Column 4 of Table C6 reports results when impact is instead measured as the *number* of high-impact articles. The coefficients on the key variables of interest are statistically indistinguishable from those obtained using the share-based measure. Notably, the marginal effect of publication quantity declines, while the marginal effect of publication impact increases substantially. This

likely reflects that the count-based impact measure partially captures non-linearities in publication quantity. Since the results are not driven by this modeling choice – and because the share measure is easier to interpret in combination with a quantity measure – we continue to use the share of high-impact publications in the main analysis.

C.5 Robustness to alternative measures of evaluator quality

We construct a single measure of evaluator quality, defined as the number of high-impact articles. An article is considered high-impact if it is published in a journal that falls within the top quartile of the Web of Science rankings for STEM+M fields or in an A-list journal designated by the ministry for Social Sciences and Humanities. We then test the robustness of our results to alternative definitions of evaluator quality. In addition, we address the concern that the mean may conceal heterogeneity in individual influence: a committee composed entirely of average researchers and one combining very high- and very low-productivity evaluators may have the same mean quality, even though their internal dynamics could differ. Highly productive members, for instance, may set the tone for evaluations and shape criteria. Analogously to the section on candidate profiles, we define the following alternative 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 for 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 indexed in the Web of Science;
4. *Quantity*: total number of articles published, without adjusting for journal quality;
5. *Quality/Impact (Citation-based)*: total number of citations accumulated by the evaluator.

Table C7 reports the results. Across all specifications, the coefficient 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 in all specifications. The interaction between *Committee Quality* and *Candidate's Publication Impact* is positive in all specifications, although

statistically insignificant when evaluator quality is proxied by *Articles in Web of Science* or *Citations*.

Table C1: Candidate research profiles in curricula vitae and OpenAlex+IRIS database

	Mean	SD
Total Publications (CV)	47.16	42.38
Total Publications (OA+IRIS)	28.02	35.14
Articles (CV)	25.97	30.17
Articles (OA+IRIS)	25.19	31.75
High-Impact Articles (CV)	10.68	16.45
High-Impact Articles (OA+IRIS)	10.75	17.09
Books (CV)	1.89	3.36
Books (OA+IRIS)	0.17	1.06
Chapters (CV)	5.87	9.19
Chapters (OA+IRIS)	0.46	1.95
Number of Candidates	46329	
Number of Candidates Matched in OpenAlex and IRIS	42472	

Notes: The sample is all candidates who participated in the first wave of the 2012 cycle of evaluations and are matched to OpenAlex and IRIS database. Productivity measures are accumulated ten years prior to the exam. *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.

Table C2: Candidate research profiles in curricula vitae and OpenAlex+IRIS database

	CV Measures				
	Total Publications	Articles	High-Impact Articles	Books	Chapters
Total Publications	.67	.	.	.	.
Articles	.	.8	.	.	.
High-Impact Articles	.	.	.81	.	.
Books	.	.	.	.05	.
Chapters	.	.	.	.	.06

Notes: The sample is all candidates who participated in the first wave of the 2012 cycle of evaluations and are matched to OpenAlex and IRIS database. *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.

Table C3: Correlation between dimensions of candidate research profiles

	Quantity	Journal Impact	AIS	Citations	Collaborators per Paper	Seniority	Topic Dispersion	FirstLastSingle Author per Paper	Novelty
Quantity	1	-.09	.32	.13	.06	-.1	-.2	-.05	.02
Journal Impact	-.09	1	.29	.13	.14	.1	-.08	-.1	.15
AIS	.32	.29	1	.3	.27	.02	-.15	-.15	.1
Citations	.13	.13	.3	1	.1	-.09	-.28	-.06	.11
Collaborators per Paper	.06	.14	.27	.1	1	.09	-.05	-.5	.14
Seniority	-.1	.1	.02	-.09	.09	1	.05	-.06	.02
Topic Dispersion	-.2	-.08	-.15	-.28	-.05	.05	1	.03	-.05
FirstLastSingle Author per Paper	-.05	-.1	-.15	-.06	-.5	-.06	.03	1	-.1
Novelty	.02	.15	.1	.11	.14	.02	-.05	-.1	1

Notes: The sample is all candidates who applied for qualification in the first wave of the 2012 cycle of the Italian qualification system. For correlations with the *Novelty* measure, we restrict the sample to STEM fields, due to its unreliability in SSH. All candidate productivity measures are standardized within each exam.

Table C4: Correlation between different dimensions of candidate research profiles and qualification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Quantity	0.113***									0.077***	0.096***
	(0.005)									(0.008)	(0.005)
Impact: Journal Impact		0.082***								0.052***	0.080***
		(0.005)								(0.005)	(0.004)
Impact: Article Influence Score			0.095***							0.064***	0.037***
			(0.008)							(0.012)	(0.007)
Impact: Citations				0.062***						0.033***	0.016***
				(0.005)						(0.005)	(0.003)
Collaborators per Paper					0.001					-0.032***	-0.024***
					(0.004)					(0.005)	(0.004)
Seniority						-0.004***				-0.000	-0.003***
						(0.001)				(0.001)	(0.001)
Topic Dispersion							-0.091***			-0.068***	-0.052***
							(0.003)			(0.006)	(0.003)
First+Last+Single Author per Paper								0.001		0.015***	0.008**
								(0.004)		(0.005)	(0.003)
Novelty									0.006	-0.010***	
									(0.004)	(0.004)	
Observations	69020	69020	69020	69020	69020	69020	69020	69020	34596	34596	69020
Adjusted R^2	0.055	0.028	0.038	0.017	-0.000	0.004	0.029	-0.000	0.000	0.132	0.113

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity measures are standardized within each exam. In columns 1-8 and 11, the sample is all candidates who applied for qualification in the first wave of the 2012 cycle of evaluations. In columns 9 and 10 the sample is all candidates who applied for qualification in STEM+M fields in the first wave of the 2012 cycle of evaluations. Standard errors are clustered at the exam level. *** p < .01, ** p < .05, * p < .1

Table C5: Committee quality and returns to different dimensions of candidate research profiles

	STEM	Overall
	(1)	(2)
Quantity	0.070***	0.074***
	(0.007)	(0.007)
Impact: Journal Impact	0.051***	0.055***
	(0.005)	(0.005)
Impact: Article Influence Score	0.072***	0.074***
	(0.011)	(0.010)
Impact: Citations	0.036***	0.031***
	(0.005)	(0.004)
Collaborators per Paper	-0.028***	-0.027***
	(0.005)	(0.005)
Seniority	0.003***	0.002**
	(0.001)	(0.001)
Topic Dispersion	-0.058***	-0.040***
	(0.005)	(0.004)
First+Last+Single Author per Paper	0.022***	0.018***
	(0.004)	(0.004)
Novelty	-0.007**	
	(0.003)	
Quantity × Committee Quality	0.007	0.008
	(0.009)	(0.009)
Impact: Journal Impact × Committee Quality	-0.028*	-0.033*
	(0.017)	(0.017)
Impact: Article Influence Score × Committee Quality	0.021	0.028
	(0.021)	(0.022)
Impact: Citations × Committee Quality	0.013	0.007
	(0.009)	(0.008)
Collaborators per Paper × Committee Quality	0.015	0.013
	(0.011)	(0.012)
Seniority × Committee Quality	0.002	0.002
	(0.002)	(0.001)
Topic Dispersion × Committee Quality	0.017	0.011
	(0.012)	(0.008)
FirstLastSingle Author per Paper × Committee Quality	0.013	0.012
	(0.010)	(0.010)
Observations	34596	41395
Number of Exams	109	184
Adjusted R^2	0.18	0.18
Candidate Connections	✓	✓

Notes: The dependent variable is a binary indicator equal to one if the candidate qualifies. All candidate productivity measures are standardized within each exam. In column 1, we restrict the sample to candidates who applied for qualification in STEM+M fields in the first wave of the 2012 cycle of evaluations. In column 2, the sample is all candidates who applied for qualification in the first wave of the 2012 cycle of evaluations. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table C6: Effect of committee composition on evaluation outcomes (alternative data sources and measures)

	Main-Text	OpenAlex IRIS	OpenAlex IRIS - Scaled	Count High-Impact
	(1)	(2)	(3)	(4)
Candidate Quantity Measure	0.113***	0.108***	0.103***	0.055***
	(0.004)	(0.004)	(0.004)	(0.006)
Candidate Impact Measure	0.094***	0.065***	0.070***	0.113***
	(0.005)	(0.004)	(0.004)	(0.008)
Committee Quality	-0.065**	-0.065**	-0.065**	-0.065**
	(0.031)	(0.031)	(0.031)	(0.031)
Candidate Impact Measure x Committee Quality	0.034***	0.034***	0.033***	0.041**
	(0.012)	(0.010)	(0.010)	(0.016)
Candidate Quantity Measure x Committee Quality	-0.023**	-0.011	-0.011	-0.042***
	(0.011)	(0.009)	(0.008)	(0.013)
Observations	69020	69020	69020	69020
Adjusted R^2	0.138	0.114	0.110	0.146

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.
*** p < .01, ** p < .05, * p < .1

Table C7: Impact of committee composition on evaluation outcomes (alternative measures of committee quality)

	Main-Text	Top Researcher	AIS	Articles WOS	Articles	Citations
	(1)	(2)	(3)	(4)	(5)	(6)
Candidate's Number of Publications	0.113***	0.046***	0.113***	0.112***	0.113***	0.113***
	(0.004)	(0.016)	(0.004)	(0.004)	(0.004)	(0.004)
Candidate's Impact of Publications	0.094***	0.071***	0.094***	0.094***	0.093***	0.093***
	(0.005)	(0.026)	(0.005)	(0.005)	(0.005)	(0.005)
Committee Quality	-0.065**	-0.100*	-0.066**	-0.061**	-0.057**	-0.017
	(0.031)	(0.057)	(0.027)	(0.029)	(0.027)	(0.034)
Candidate's Number of Publications x Committee Quality	-0.023**	-0.025	-0.014	-0.016	-0.025**	-0.017**
	(0.011)	(0.021)	(0.011)	(0.011)	(0.011)	(0.008)
Candidate's Impact of Publications x Committee Quality	0.034***	0.037	0.026**	0.023*	0.029*	0.012
	(0.012)	(0.026)	(0.012)	(0.013)	(0.015)	(0.014)
Observations	69020	69020	69020	69020	69020	69020
Adjusted R^2	0.138	0.137	0.138	0.137	0.137	0.135

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. *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. 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.
*** p < .01, ** p < .05, * p < .1

D The marginal treatment effect of qualification

We first estimate the effect of qualification using variation in candidates' connections to the randomly drawn committee. Candidates differ in whether they have strong professional connections to the evaluators selected for their field, so this instrument varies across candidates even within the same examination. We construct the shock to strong connections by subtracting the number expected given the eligible evaluator pool and the institutional rules governing the draw. We then use this residualized connection measure as an instrument for qualification in a 2SLS specification. The resulting estimates identify a local average treatment effect for candidates whose qualification decisions are shifted by unexpectedly strong committee connections.

Table B1 shows that the connection instruments strongly predict qualification. Table B2 reports the second-stage estimates. Qualification increases promotion within five years by 47 percentage points among Associate Professor candidates, consistent with relaxing a formal barrier to short-run advancement. By contrast, it has no detectable effect on subsequent publications, high-impact articles, citations, or promotion at longer horizons.

We next estimate marginal treatment effects following Heckman and Vytlacil (2007) and Carneiro et al. (2011). Let $D_i \in \{0, 1\}$ denote qualification, Y_i an outcome, X_i observed candidate characteristics, and Z_i the excluded connection instruments. Qualification follows the latent-index model

$$D_i = \mathbf{1}(P(Z_i, X_i) \geq U_i), \tag{B1}$$

where $P(Z_i, X_i) = \Pr(D_i = 1 \mid Z_i, X_i)$ is the propensity score and $U_i \sim \text{Unif}[0, 1]$ measures unobserved resistance to qualification. Candidates with lower values of U_i are more likely to qualify.

Let Y_{1i} and Y_{0i} denote potential outcomes with and without qualification. The marginal treatment effect is

$$\text{MTE}(x, u) = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x, U_i = u]. \tag{B2}$$

It measures the gain from qualification for candidates with observed characteristics x and unobserved resistance u.

Identification relies on variation in Z_i, which captures connections to members of the evaluation committee. It requires:

1. *Independence* is satisfied due to the random nature of the draw. Conditional on the assignment

mechanism, we identify exogenous variation in connections following Borusyak and Hull (2023).

$$(Y_{1i}, Y_{0i}, U_i) \perp Z_i \mid X_i \tag{B3}$$

2. *Relevance* requires that the excluded shifters substantially affect the probability of qualification. Table B1 reports the test of joint significance of the instruments. The chi-squared test statistic is ≈ 70.

$$\Pr(D_i = 1 \mid Z_i, X_i) \tag{B4}$$

3. *Support* requires that the propensity score has support on an interval $\mathcal{P} \subseteq (0, 1)$ with sufficient variation induced by Z_i. This ensures that candidates can be observed at different points along the qualification margin.

Under these assumptions, the marginal treatment effect is identified as

$$\text{MTE}(x, u) = \frac{\partial \mathbb{E}[Y_i \mid X_i = x, P(Z_i, X_i) = p]}{\partial p}\bigg|_{p=u}. \tag{B5}$$

Intuitively, random variation in the number of connections to the committee shifts candidates along the qualification margin by changing their probability of qualification. The derivative of the outcome equation with respect to the propensity score traces out the return to qualification for candidates at different points of this margin.

We estimate the propensity score using a logit model with candidate characteristics, examination fixed effects, and connection variables. The estimates confirm that connections significantly affect the probability of qualification and generate substantial variation in the propensity score (Table B1). Figure B1 shows substantial overlap between qualified and rejected candidates over $p \in [0.1, 0.8]$. The outcome equation includes second-order polynomials in the propensity score and candidate characteristics, together with their interactions. We evaluate the MTE at 30 propensity-score values, holding observables at their means, and bootstrap both estimation stages.

Figure B2 reports the MTE of qualification on future citations. The estimated effect is negative among candidates with low latent resistance and becomes less negative as resistance rises. This pattern may reflect mean reversion among candidates with unusually high pre-exam output or a shift toward administrative and service responsibilities following qualification. Importantly, the

effect is not significantly positive at any point of the observed margin.

These estimates help distinguish better selection by better-published committees from heterogeneous effects of qualification. The lower future citations of candidates they reject cannot be explained by denying qualification to candidates who would have benefited substantially: qualification has, if anything, a negative citation effect for marginal candidates. Similarly, the higher citations of candidates they qualify are unlikely to reflect higher returns to qualification, since the MTE is never significantly positive. The citation results therefore support the interpretation that better-published committees rank candidates more accurately rather than selecting candidates with different treatment effects.

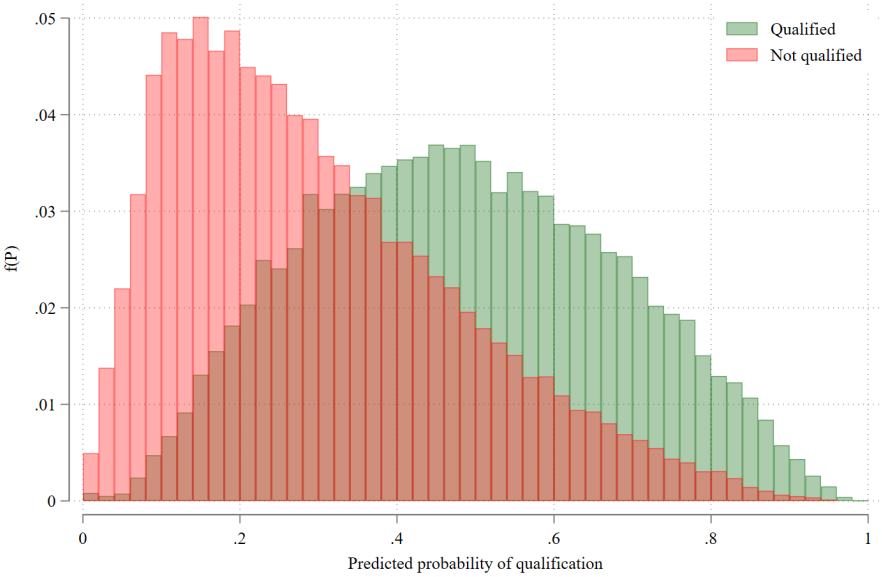
Thus, the MTE estimates on future citations support the interpretation that better-published committees improve selection quality by ranking candidates more accurately.

The career outcomes show a different pattern. Figure B3 indicates that qualification accelerates advancement among Associate Professor candidates with low latent resistance, both into Associate Professor positions (Panel A) and subsequently to Full Professor (Panel B). These effects decline with latent resistance, suggesting that qualification matters most for candidates already close to promotion.

Because qualification itself facilitates career advancement, the higher promotion rates of candidates selected by better-published committees may reflect both stronger underlying prospects and causal effects of qualification. Career outcomes are therefore less diagnostic than citations.

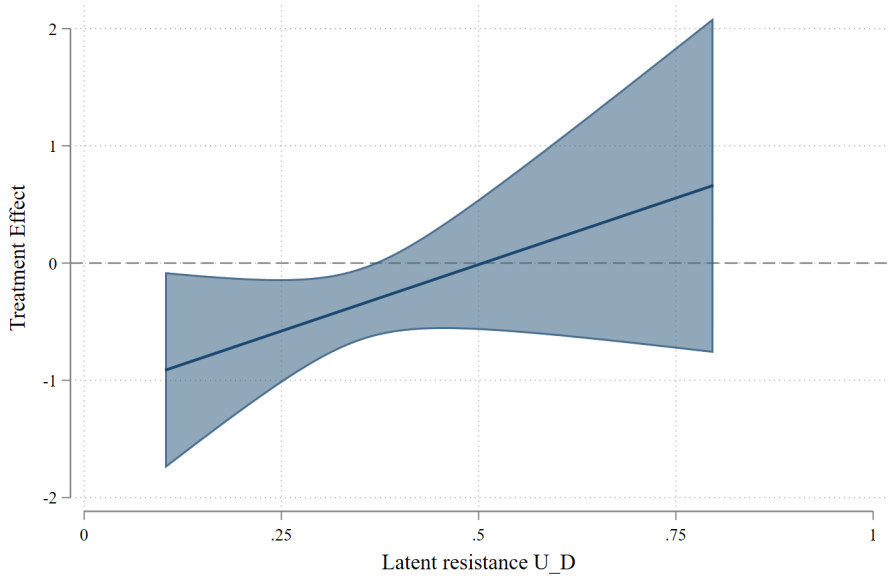
Overall, the MTE evidence strengthens the interpretation of the citation results as improved selection on underlying research potential. It provides less definitive evidence for career progression, where qualification itself affects subsequent advancement.

FIGURE B1: SUPPORT OF ESTIMATED SUCCESS PROBABILITY BY QUALIFICATION STATUS



Notes: The figure displays the support of the estimated propensity score for qualified and non-qualified candidates. The propensity score is obtained from the logit model reported in Table B1. Each bar represents the fraction of observations within a given bin of the propensity score.

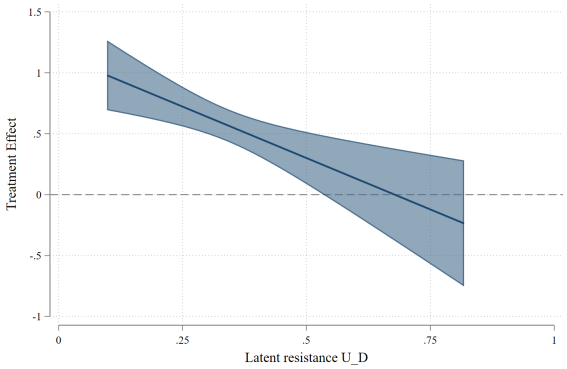
FIGURE B2: MARGINAL TREATMENT EFFECT OF QUALIFICATION ON FUTURE RESEARCH OUTPUT



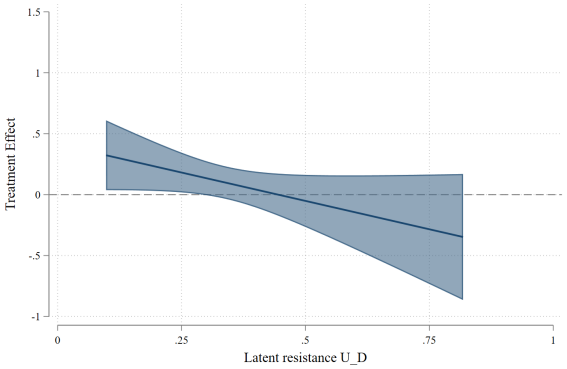
Notes: The figure reports the marginal treatment effect of qualification on the cumulative number of future citations. The measure is normalized within examinations. To estimate the function plotted here, we first estimate a partially linear specification of the promotion outcome on flexible functions of the predicted probability of qualification, $g(P)$, a second order polynomial in covariates $f(X)$, and interactions between P and $f(X)$. The MTE curve is obtained by evaluating the derivative of the estimated conditional expectation function with respect to P at the average value of X. The solid line reports the estimated MTE; the shaded area reports the 90 percent bootstrap confidence interval (250 replications).

Figure B3: Marginal Treatment Effect of Associate Professor Qualification on Promotion

Panel A.
Promoted to Associate Prof. within 5 years



Panel B.
Promoted to Full Prof. within 10 years



Notes: The figure reports the marginal treatment effect of qualification to Associate Professor on the probability of subsequent promotions. To estimate the function plotted here, we first estimate a partially linear specification of the promotion outcome on flexible functions of the predicted probability of qualification, $g(P)$, a second order polynomial in covariates $f(X)$, and interactions between P and $f(X)$. The MTE curve is obtained by evaluating the derivative of the estimated conditional expectation function with respect to P at the average value of X. The solid line reports the estimated MTE; the shaded area reports the 90 percent bootstrap confidence interval (250 replications).

Table B1: First-stage estimates of connections on qualification status

	Average derivative
	(1)
Panel A. Candidate observables	
Total publications	0.155***
	(0.003)
High-impact publications	0.096***
	(0.002)
Citations	0.071***
	(0.003)
Panel B. Exogenous shifters	
Coauthor	0.195***
	(0.035)
Same affiliation	0.160***
	(0.031)
Observations	69,015
Chi-squared for test of excluded instruments	72.04

Notes: The table reports average marginal derivatives from a logit model of qualification on candidates' observable characteristics and their connections to the committee. The sample is all candidates who submitted an application in the first wave of the 2012 cycle of the Italian evaluation system. All candidate productivity measures are normalized within each exam. The specification includes exam fixed effects. Standard errors are clustered at the exam level. The chi-squared test reports the joint significance of exogenous shifters.
*** p < .01, ** p < .05, * p < .1

Table B2: Impact of qualification on future outcomes

	(1) Total Publications	(2) High-Impact Articles	(3) Citations (All Articles)	(4) Citations (Pre-Exam Articles)	(5) Citations (Post-Exam Articles)	(6) Promoted Associate Prof. (within 5 years)	(7) Promoted Associate Prof. (within 10 years)	(8) Promoted Full Prof. (within 5 years)	(9) Promoted Full Prof. (within 10 years)
Panel A. Associate Professor Candidates									
Qualified	0.136	-0.087	-0.108	-0.267	0.120	0.466***	0.181	0.005	0.004
	(0.305)	(0.330)	(0.319)	(0.244)	(0.359)	(0.149)	(0.152)	(0.029)	(0.121)
Observations	47424	47424	47424	47424	47424	47424	47424	47424	47424
R^2	0.17	0.15	0.34	0.59	0.14	0.14	0.08	0.01	0.03
Panel B. Full Professor Candidates									
Qualified	0.227	0.326	0.373	-0.171	0.654			0.155	-0.105
	(0.426)	(0.431)	(0.431)	(0.266)	(0.528)			(0.165)	(0.215)
Observations	21589	21589	21589	21589	21589			21589	21589
R^2	0.25	0.22	0.41	0.64	0.13			0.11	-0.00

Notes: The table reports 2SLS estimates of qualification on future outcomes. Qualification is instrumented by the deviation of realized connections from its expected value under the assignment mechanism. The dependent variables measure candidate outcomes ten years after the evaluation unless otherwise stated. In Panel A. the sample is candidates who submitted an application to Associate Professor in the first wave of the 2012 cycle of the Italian evaluation system. In Panels B. the sample to candidates to Full Professor. *Promoted Associate Prof.* and *Promoted Full Prof.* are dummy variables indicating whether the candidate is promoted to Associate and Full Professor, respectively. Candidate outcomes are normalized at the examination level. We include four dummy variables for broad disciplinary group, control for candidates' number and impact of publications, and the number of pre-exam citations. We also include indicators for candidates' affiliation and academic rank at the time of the evaluation. Standard errors are clustered at the examination level.

*** $p < .01$, ** $p < .05$, * $p < .1$