Top Researchers on Scientific Committees: Decision Outcomes, Peer Dynamics, and Opportunity Costs

Milan Makany Natalia Zinovyeva[*]

January 21, 2026

Abstract

Science disproportionately relies on top researchers to evaluate the work of others, potentially diverting their scarce time from producing new knowledge. Using random assignment of evaluators in Italy's national academic promotion system, we show that committees with better-published members select candidates who subsequently receive more citations and experience faster career advancement. Better-published evaluators also change committee dynamics: they raise peers' effort and induce convergence toward more impact-focused evaluation criteria, consistent with reputational pressures. Committee service, however, carries opportunity costs. Serving on a two-year committee reduces evaluators' own publication output by about 20% of a typical year's production, with particularly large effects for highly productive researchers working in small teams. These findings suggest that evaluation systems that account for peer effects and heterogeneous opportunity costs – through committee design and service allocation – can improve evaluation efficiency while reducing the overall burden on science.

JEL Codes: I23, J45, M51, D71, D73

Keywords: committees, academic promotions, experts, evaluation 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, 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 Edinburgh; University of Warwick; Erasmus School of Economics; University of Ottawa; Aalto University; SOFI; and CEBI at the University of Copenhagen.

1 Introduction

Committees are central to shaping outcomes across a wide range of institutions, from corporate and professional licensing bodies to scientific funding agencies, journal editorial boards, and academic promotion panels. Across these settings, especially in academia, the demand for committee service increasingly exceeds the supply of qualified evaluators. Journals report sending more invitations per completed review and speak of "reviewer fatigue"; funders document persistent difficulty in staffing grant panels; publishers and scholarly societies experience difficulty recruiting and retaining editorial board members.[1] Evaluation service entails substantial costs: the journal peer-review system alone may absorb more than 100 million researcher-hours annually.[2] A large share of this burden falls on the most qualified experts, who become particularly difficult to recruit for evaluation service.[3]

This naturally raises the question of how much, and through which margins, do more qualified evaluators improve committee decisions, and at what cost? While it is often presumed that attracting more accomplished experts improves the quality of decision-making, theory yields ambiguous predictions. At the individual level, such experts may provide more accurate signals of quality, but they also face higher opportunity costs of time, which can discourage effort or participation (Holmstrom and Milgrom, 1991). At the group level, the presence of accomplished peers can strengthen within-committee reputational motives, encouraging others to exert effort and demonstrate competence (Swank and Visser, 2023). External reputational concerns may instead weaken incentives: when committees include a more merited member, outsiders may attribute decision quality primarily to this expert, lowering the perceived return to effort for other members (Visser and Swank, 2007). Finally, when some members are better informed, others may optimally reduce their own effort, leading to potentially lower average information acquisition (Feddersen and Pesendorfer, 1996; Persico, 2004; Chan, 2021).

Despite this rich theoretical foundation, there is little causal evidence on whether these mech-

[1] The *Nature* coverage of the Publons Global State of Peer Review (survey of 11k+ researchers) describes editors "struggling to find willing reviewers" and a rise in invitations per completed review (Publons, 2018; Vesper, 2018). Funders have acknowledged recruitment challenges and weak incentives; see the UKRI review of interventions to optimize grant peer review (Kolarz et al., 2023). Surveys note that time burden is a major factor behind recruitment and retention challenges for editorial board roles (Funk et al., 2024).

[2] Aczel et al. (2021) estimate over 100 million reviewer hours in 2020 across major countries, with implied costs exceeding $2.5 billion if expressed in average hourly wages.

[3] Publons report highlights a heavy concentration of workload, with 10% of reviewers being responsible for 50% of peer reviews (Publons, 2018; Vesper, 2018). This is echoed by first-person accounts on *Nature Careers* of cutting back on reviewing amid heavy demand (Dance, 2023).

anisms operate in practice. We do not know whether better experts on committees apply different criteria and whether more accomplished researchers ultimately improve the quality of committee decisions. It is also unclear whether other committee members adjust their effort in response to the presence of more accomplished researchers, as assumed in models of strategic information acquisition and reputation. Finally, we lack evidence on whether the core premise of multitasking models – that high-ability individuals face steeper opportunity costs that shape their allocation of effort – is empirically relevant for committee service. Providing evidence on these mechanisms is important both to assess the empirical relevance of existing theories and to guide the design of committees – including their composition, procedures, and evaluators' incentive systems.

In this paper, we exploit the random assignment of evaluators to identify the causal effect of evaluator expertise on committee dynamics and assessment outcomes. Our empirical setting is Italy's *Abilitazione Scientifica Nazionale (ASN)*, a centralized promotion system overseeing all Italian public university promotions since 2012. Candidates for associate and full professor positions must first obtain a qualification from discipline-level committees before applying for promotion at the university.[4] Each committee consists of five evaluators, randomly drawn from a pool of eligible professors who volunteer, and serve a two-year term. Our database covers around 10,000 professors eligible to serve, of whom about 3,200 sat on more than 680 committees between 2012 and 2021. It also includes all applications from the first national round of evaluations (around 60,000) and roughly 300,000 evaluation reports written by evaluators for individual candidates. Since evaluators are randomly assigned, committees differ exogenously in the expertise of their members: some, by chance, include more highly productive researchers, whom we refer to as "expert" evaluators.

We exploit this random variation in committee composition to address four questions: (i) whether better-published committees differ from others in their evaluation criteria; (ii) whether evaluator expertise improves the quality of committee decisions, as reflected in the subsequent performance of selected candidates; (iii) whether better experts generate peer responses within committees in both effort provision and evaluation criteria, and whether evaluators' own effort varies with expertise; and (iv) whether more accomplished researchers face higher opportunity costs of service – measured by foregone research output – and whether these costs reduce their willingness to participate in the future.

We find that committees with better-published experts apply stricter standards and place

[4] There are 184 disciplinary areas, grouped into 14 broad fields. Committees are formed at the level of these narrowly defined areas. For instance, within economics there are several committees, including economic history (*storia economica*), econometrics (*econometria*), and finance (*scienza delle finanze*).

greater emphasis on publication impact over quantity. Replacing an average evaluator with one whose pre-committee publication record is one standard deviation stronger lowers a candidate's probability of success by 1.5 percentage points (4% of the 37% baseline success rate). At the same time, the increase in evaluator quality raises the marginal effect of publication impact on the candidate's success and decreases the contribution of publication quantity to candidate's success.

The differences in assessment criteria across committees with different publication quality translate into meaningful improvements in the quality of selection. Candidates approved by better-published committees subsequently produce similar number of publications but place a higher share of publications in top journals and receive significantly more citations, both for their prior work and for the work produced after evaluation. Selected candidates are also more likely to achieve subsequent promotions. Among those qualifying for Associate Professor, replacing an average evaluator with one whose pre-committee publication record is one standard deviation stronger increases the probability that selected candidates become full professors within the next ten years by about one percentage point (a 6% change relative to the 17% baseline). These differences persist even after controlling for a wide set of candidate characteristics, including past research productivity and affiliation. Since the composition of national evaluation committees is highly unlikely to directly affect candidates' later careers beyond the qualification outcome itself, these results suggest that committees with better-published researchers are better able to identify talent and recognize candidates' "latent" potential.

We then use data on individual-level voting to test whether better experts shape committee decisions only through their own evaluations or also through their influence on peers. We show that better-published experts are not only themselves less likely to cast positive votes and place greater weight on publication impact over quantity, but they also make their peers adopt similar standards. Their peers also devote greater effort to their reports than when they serve on committees with less merited colleagues. Specifically, when an average evaluator is placed on a committee where a peer's average publication record is one standard deviation stronger than expected, she increases the length and lexical complexity of her evaluation reports by about 7% of a standard deviation. We find no evidence that better-published evaluators themselves exert less effort in the evaluation task. On average, their reports are similar in length and linguistic complexity to those of other evaluators.

While expert involvement enhances the quality of evaluation outcomes, it also entails substantial productivity losses for researchers. Serving on an ASN committee requires evaluating hundreds of candidates over a two-year term, in addition to regular academic responsibilities.[5] We find that random assignment to a committee leads to a substantial decline in researchers' publication output during and after the service term. Using comprehensive bibliometric data on all potential evaluators, we estimate that committee service reduces output by roughly 20% of a typical year's publications, with the effect spread over several years.[6]

The magnitude of this productivity loss depends crucially on the size of researchers' collaboration networks, with those working in smaller teams experiencing substantially larger declines. At the same time, top researchers working in smaller teams bear the steepest costs. These patterns indicate that collaborators can partially absorb the disruption caused by committee service, but that the opportunity cost of service remains highest for the most productive researchers when such buffers are absent.

The costs of committee service extend beyond research output. Professors drawn to committees subsequently supervise fewer PhD graduates, with effects emerging about four years after the draw – consistent with a reduction in the intake of new students rather than delayed completion. As with research productivity, these effects are attenuated for researchers embedded in larger teams, suggesting that collaboration helps absorb time shocks. By contrast, we find no systematic differences by pre-service research quality in supervision outcomes.

Although better-published evaluators do not appear to exert less effort while serving, they are substantially less likely to volunteer for future committees, consistent with a higher opportunity cost of their time. A one–standard deviation increase in pre-committee publication quality corresponds to a 3.4-percentage-point lower probability of volunteering again in subsequent waves – an 18% decrease relative to the 19% baseline.

Aggregate participation patterns mirror this individual-level response: over time, top researchers have become markedly less likely to serve. This selective attrition risks skewing the evaluator pool away from the most qualified researchers – the very individuals who contribute most to evaluation quality but who face the highest personal costs of participation.

This paper makes several contributions to the literature. First, we extend research on how evaluator characteristics shape committee decisions. While prior work has examined the roles of gender

[5] Informal discussions with evaluators across fields and waves suggest a workload of roughly 200–300 hours over a two-year term. Between 2012 and 2018, each committee evaluated on average about 270 candidates (Italian National University Council, 2023).

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

(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' expertise 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 and shift the emphasis from publication quantity to impact. This change in evaluation criteria leads committees to select candidates who produce more impactful subsequent research and experience faster career progression, demonstrating that better expert involvement improves decision quality. Our results thus 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.

Second, we contribute to the literature on committee decision-making and peer effects in professional settings. We show that the presence of highly qualified experts systematically alters the behavior of other committee members: peers devote more effort, apply stricter standards, and place greater emphasis on research impact when serving alongside top researchers. This pattern is consistent with the hypothesis that within-committee reputation concerns influence behavior (Swank and Visser, 2023). In this theoretical framework, members exert greater effort when paired with high-ability peers to maintain their reputation in the eyes of those peers. Consistent with this mechanism, we find that evaluators increase effort, rather than reduce it, when serving with more accomplished experts. We thus provide the first causal evidence supporting this hypothesis in a field setting. More broadly, our findings validate ideas from a rich theoretical literature on group decision-making, which argues that in information-sharing committees the presence of an informed member can change others' incentives (Ottaviani and Sørensen, 2001), and that career (Levy, 2007) and reputation concerns (Prat, 2005; Visser and Swank, 2007) can alter behavior in such settings. Notably, our results suggest that even in an environment of full transparency — where individual evaluations and votes are made public — within-committee reputation remains a dominant force in shaping behavior. Our findings also complement evidence from personnel economics showing that coworkers' productivity affects individual effort (e.g., Chan 2021 in medical teams). We document large peer effects in a high-stakes, professional context: committee members work harder and apply different evaluation standards when seated alongside an accomplished expert, thereby improving the quality of collective decisions.

Third, we advance understanding of the opportunity costs of academic service. It is widely recognized that activities such as peer review and committee work demand substantial, and often uncompensated, time from researchers. However, prior evidence on the impact of these duties was based mainly on self-reported time use. We provide the first causal evidence on how evaluation service affects a researcher's own productivity, documenting a large and persistent decline in research output following service. Importantly, we show that this cost is strongly mediated by researchers' collaboration networks: larger teams partially absorb the time shock induced by service, while researchers working in smaller teams experience substantially larger losses. Even top researchers – who might be expected to multitask more efficiently – cannot fully shield their research output when such buffers are absent. Consistent with these opportunity costs, top researchers also become less likely to volunteer for evaluation service over time. These findings have implications for incentive design in academia. They underscore a core insight from multi-tasking theory (Holmstrom and Milgrom, 1991): when socially valuable tasks such as evaluation carry high personal costs and limited rewards, agents under-invest in them. Our results further contribute 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) by highlighting the role of team structure in shaping researchers' ability to absorb service-related time demands.

Our findings raise concerns about the sustainability of centralized evaluation systems. Promotion and grant review processes depend on the participation of highly qualified scholars, yet the opportunity cost of service may lead precisely these researchers to withdraw over time. As a result, evaluation standards may erode not through formal rule changes, but through gradual shifts in who is willing to serve. This mechanism is particularly salient in debates over the role of peer review versus bibliometric indicators (DORA, 2012; Hicks et al., 2015; Bertocchi et al., 2015; Stephan et al., 2017; Heckman and Moktan, 2020): while expert judgment improves selection, its effectiveness ultimately depends on sustained expert participation. The challenge is therefore not choosing between metrics and peer review, but ensuring that evaluation systems are designed to make expert service viable in the long run – a concern that extends beyond academia to professional licensing, regulatory commissions, and other decision-making bodies.

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 committee outcomes, peer effects among evaluators, and the costs to evaluators. Section 5 concludes with implications for policy and theory.

2 Italian National Academic Qualifications

Italy's *Abilitazione Scientifica Nazionale* (ASN) is a nationwide qualification process introduced in 2012 as part of the Gelmini Reform (Law 240/2010). The ASN acts as a mandatory first-stage filter for academic promotions. Candidates seeking promotion to Associate or Full Professor must first obtain a national qualification (*abilitazione*) from a discipline-specific committee. Only candidates who obtain this qualification are eligible to apply for positions or promotions at the university level. This two-stage system was designed to homogenize evaluation standards across universities and limit local favoritism.

2.1 Evaluator selection

Evaluation committees are formed at the disciplinary field level and consist of five members.[7] Committee members are selected by random draw from a pool of eligible professors who volunteer for service and satisfy field-specific research productivity thresholds. Eligibility criteria reflect disciplinary publication norms. In STEM fields (including medicine) and psychology, eligibility is based on minimum thresholds for journal publications, citations, and the H-index. In Social Sciences and Humanities (SSH), criteria include minimum numbers of journal articles, publications in high-impact journals, and books or book chapters. These thresholds determine the pool of professors eligible to be drawn within each field.

The draw of committee members is implemented centrally by the Ministry of Education using a computer-generated randomization procedure. Each eligible professor is assigned an identifier, and a random sequence determines the order in which names are drawn. Committees are filled sequentially, subject to two institutional constraints: (i) no more than one evaluator may come from the same university, and (ii) in large fields, sub-field quotas ensure representation of major areas of specialization.[8]

Conditional on these constraints, assignment is random. There is no scope for strategic manipulation or candidate influence in the selection of committee members. In expectation, realized committee composition mirrors the characteristics of the eligible pool within each field, and deviations from expected composition arise solely from the random component of the draw. This source

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

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

of variation underlies our empirical strategy.

In the first ASN round (2012–13), four committee members were drawn from Italian universities and one from a list of eligible Italian scholars working abroad in OECD countries. This foreign-member requirement was dropped after 2013, and in all subsequent rounds all five members were drawn from Italy-based faculty.[9]

If a selected evaluator resigns, a replacement is drawn following the original randomization procedure. To address the possibility that resignations are not random, our empirical strategy instruments final committee composition with the composition initially drawn.

Committee members serve for approximately two years and typically evaluate multiple consecutive application rounds. For example, committees appointed in late 2012 evaluated candidates who applied in 2012 and again in 2013. In the first cohort, candidates applied before committee composition was known; in subsequent cohorts, candidates observed committee membership at the time of application. After completing a term, evaluators are ineligible to serve again for at least three years. Service is essentially uncompensated: committee members receive no remuneration beyond travel per diems, and teaching-load reductions, which evaluators could in principle apply for, were minimal or rarely implemented in practice. Committee service therefore constitutes a largely pro bono academic duty.

2.2 Evaluation procedure

Evaluations are based exclusively on candidates' CVs and up to ten submitted publications. Committees meet to define evaluation criteria and assess applications, with the number and format of meetings depending on application volume and evaluation cycle. In the 2012 cycle, committees had full discretion in setting standards, although the Ministry recommended reference to three bibliometric indicators – the same indicators as used to establish evaluator eligibility. In later cycles, minimum bibliometric conditions were formally introduced, while committees retained discretion over additional criteria.

Decisions are made by qualified majority, requiring at least four out of five votes in favor. Each evaluator must provide a written report accompanying every individual vote, and the committee jointly produces a summary document for each candidate. Following the evaluation, candidates' names, evaluators' identities, individual reports, and final decisions are published online.

[9] Contemporary commentary noted practical and legal difficulties associated with the foreign-member requirement, including higher resignation rates, uncertainty about equivalence of academic ranks, and language barriers in some fields (e.g. ROARS, 2014).

The production of individual evaluation reports is a central component of committee work. Each evaluator writes an assessment for every candidate, and committee meetings largely revolve around reviewing, uploading, and aggregating these reports into collective decisions. Report writing is therefore not ancillary but a core component of evaluators' service.

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.[10] The resulting panel covers approximately 90,000 professors across 95 public universities and reports affiliation, department, research area, academic rank, and gender.

National Qualification Evaluations (ASN)

We use administrative data on all ASN participants.[11] 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 all four ASN evaluation cycles conducted between 2012 and 2021.

For the 2012 cycle, we further use the database assembled contemporaneously by Bagues et al. (2017) from the Ministry website.[12] These materials contain the names and CVs of all pre-registered candidates, as well as the individual and committee evaluation reports posted as outcomes of the

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

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

[12] Candidates' CVs, evaluation reports, and the names of unsuccessful candidates are available online for only six months after the decision.

process. The reports allow us to observe each evaluator's individual vote on each candidate and to construct text-based proxies for evaluator effort (report length and linguistic complexity).

Bibliographic databases

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

First, we extract candidates' publication records directly from the CVs submitted with ASN applications. This source mirrors the information available to evaluators at the time of decision-making, allowing us to reconstruct candidates' observable research output at evaluation.

Second, for both candidates and evaluators, we compile publication and citation histories from two external bibliometric databases. The primary source is IRIS (Institutional Research Information System), managed by ANVUR. IRIS contains publication records for researchers at Italian public universities and research centers. Publications are self-reported by researchers through an online portal, and 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.[13] 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; the full merge procedure is described in Appendix A. Together, IRIS and OpenAlex cover 96% of professors in our administrative dataset.

Doctoral graduates database

We collect data on PhD graduates to measure doctoral supervision outcomes from the National Central Library of Florence (BNCF). In Italy, doctoral theses are subject to legal deposit at the National Central Libraries of Florence and Rome.[14] In practice, the digital visibility of theses

[13] 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).

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

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 conservative and well below typical doctoral production rates at Italian universities. Applying this criterion excludes approximately 43% of institutions, which display sporadic or near-zero records inconsistent with known PhD activity and therefore likely reflect incomplete catalog coverage rather than true absence of doctoral supervision.

3.2 Descriptive statistics

Evaluators

Across the four waves of examinations, there were 685 evaluation committees in 184 fields, with around 10,000 Italy-based professors volunteering and qualifying for the pool of eligible evaluators. The average eligible pool consists of 26 professors. In the 2012 round, one of the five committee members was drawn from volunteers based outside Italy.[15] Our analysis focuses on Italy-based evaluators.[16]

Appendix Table A1 summarizes the productivity of professors in the pool of eligible evaluators during the ten years prior to the exam. Panel A. shows that the average researcher in the sample has 60 publications, about 90% of which are journal articles; the remainder includes books, chapters, conference proceedings, and other outputs. Roughly 40% of journal articles appear in high-impact outlets. We define high-impact journals using field-specific classifications. In STEM+M we use journals in the top quartile (Q1) of Web of Science by Article Influence Score (AIS).[17] In economics,

[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] Foreign evaluators wrote substantially shorter reports: on average 35 words fewer (about 20% of the 182-word average for Italian evaluators), which consistent with the idea of their weaker engagement. All results are robust to including foreign evaluators, though estimates are noisier.

[17] 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.

business, and other SSH fields, we use the Italian evaluation agency's list of "A-journals" (about 7,000 journals). This hybrid definition improves comparability across fields, since continuous impact metrics are not consistently available in SSH. On average, researcher in our sample supervised 3 students over this period.

Our baseline measure of *evaluator productivity* is the number of high-impact journal articles published in the ten years preceding the exam. This metric captures both output volume and scientific influence, aligns with evaluation practices in Italy, and has a direct policy interpretation.[18] To account for cross-field heterogeneity, we normalize this productivity measure within each evaluator pool (discipline × exam cohort).

For interpretation, we also define *top researchers* as professors in the top 25% of all Italian full professors by this metric. Panel A. of Appendix Table A3 show that eligible volunteers are systematically better published than the average full professor, reflecting eligibility requirements and self-selection into volunteering. Panel B. shows that approximately 47% of volunteers qualify as top researchers under this definition.[19]

We measure *committee quality* as the average of evaluators' normalized productivity. This measure exploits continuous variation in committee strength and maps naturally to the intensity of scrutiny when each member incrementally contributes to standards and decisions.[20]

Around 9% of initially drawn evaluators resigned and were replaced by other randomly selected eligible evaluators. Since inclusion in the pool requires volunteering in advance, resignations plausibly reflect idiosyncratic events. Consistent with this interpretation, Panel C. of Appendix Table A3 shows only marginally significant differences between evaluators who resigned and those who did not. On average, better-published researchers were more likely to resign.

Candidates

The 2012 ASN cycle was the largest, with 69,020 pre-registered applications (about 375 per field). Its exceptional scale reflected the absence of promotions during the transition from the previous system, the low cost of applying, and uncertainty about the criteria committees would adopt. More than 46,000 researchers submitted applications, corresponding to around 61% of assistant professors

[18] Compared with citation counts, journal rankings reduce differences in citation practices across subfields and lessen the risk of understating novel work that accrues citations with delay (Wang et al., 2017; Koffi, 2025).

[19] Panels B. and C. of Appendix Table A1 show that top researchers are, on average, more productive both with respect to publications and student supervision.

[20] Appendix C.5 shows robustness to alternative committee-quality measures.

and 60% of associate professors in Italy.[21] About one third of applicants registered in multiple fields (e.g., Political Economy and Applied Economics) or for both associate and full professorships within the same field.

Candidates could withdraw after committees were drawn and criteria announced. The Bagues et al. (2017) database includes CVs of all pre-registered candidates, posted online before the final list of evaluators was confirmed. Roughly 14% of candidates withdrew, leaving 59,151 applications to be evaluated (Italian National University Council, 2023).[22] Appendix Table A4 shows that withdrawals were not systematically related to assigned committee quality. Overall, 37% of pre-registered candidates (43% of evaluated candidates) were successful.

Table A5 summarizes candidate characteristics based on CVs. Two thirds of applicants were already employed at an Italian university, typically in a permanent position in the field for which they sought promotion.[23] Associate Professor applicants published their first paper 14 years before the exam and listed 52 works on their CVs on average. Full Professor applicants had 20 years since first publication and listed 87 works. About 56% of listed outputs were journal articles, and about 38% of those articles were in high-impact journals.

We use CVs to validate external publication data. In the combined OpenAlex+IRIS database we find 29 publications for the average candidate, 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 CVs and in OpenAlex+IRIS is 78%, suggesting that OpenAlex+IRIS captures journal-based output reliably. Correlations are substantially lower for books and chapters.[24]

To study how evaluator expertise shapes evaluation criteria, we focus on two widely used dimensions of candidates' research profiles: *quantity* and *impact*. We measure *quantity* as the total number of academic outputs listed on the CV (journal articles, books, chapters, conference proceedings, and patents). We measure *impact* as the share of journal articles published in top-ranked journals. Because there is no straightforward impact metric for non-journal outputs, we restrict the impact measure to journal articles and assume impact is positively correlated across output types. Defining impact as a share also avoids conflating it with nonlinear effects of quantity.[25]

[21] Based on our calculations using Ministry records on all assistant (*Ricercatori*) and associate professors (*Associati*) as of December 31, 2012.

[22] In the dataset collected by Bagues et al. (2017), this number is 59,150.

[23] Among associate professor applicants, 42% held tenured assistant professorships (*Ricercatore a tempo indeterminato*).

[24] 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.

[25] Appendix C.2 shows that these measures have the strongest predictive power for evaluation outcomes compared

Publication practices differ sharply across fields. In STEM+M, the mean share of top-quartile articles per candidate ranges from 30% to 50%. In economics and business, the mean share is about 15% for Q1 journals and 21% for A-journals; in the remaining SSH fields, it is about 29% for A-journals. We therefore normalize both quantity and impact within discipline.

When analyzing committee composition, we control for candidate–evaluator connections to avoid confounding publication-profile similarity with professional proximity. About 15% of candidates had either a co-author or a departmental colleague on their committee.[26]

For candidates employed at Italian universities, we observe both field (*settore concorsuale*) and subfield (*settore scientifico-disciplinare*) of appointment. There are 184 fields and about 370 subfields; in roughly 60% of cases, a candidate and an eligible evaluator belong to the same subfield.

Finally, we track post-ASN career outcomes using administrative records. Among candidates qualified for Associate Professor, 68% were promoted to associate professor and 17% to full professor within ten years. Among those qualified for Full Professor, 57% attained a full professorship within ten years.

Individual evaluation reports

Each committee member wrote an individual evaluation report for every application. Reports summarize the candidate's research production, provide a qualitative assessment of quality and fit, and conclude with a recommendation on qualification. The dataset contains about 295,000 reports, of which 241,744 were written by Italy-based evaluators.[27] Roughly 45% of reports express a positive vote (Appendix Table A7). Committees display substantial unanimity: about 86% of decisions are unanimous.

We use report text to construct proxies for evaluator effort. For each report, we compute length and lexical complexity. Complexity is measured using inverse document frequency (IDF). We preprocess texts by removing punctuation, numerals, stop-words, and first names, and apply stemming to account for gendered forms in Italian. We then compute $IDF_{t,i}$ for each word t in report i as the log of the inverse share of reports containing t, assigning greater weight to rarer words. We summarize complexity using *Average IDF* (mean $IDF_{t,i}$ within a report) and *Total IDF* (sum of

to citations, average AIS, number of collaborators, seniority, topic diversity, and novelty. Appendix C.4 shows robustness to alternative definitions of quantity and impact.

[26] 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.

[27] Due to a technical issue, reports are missing for 202 applications.

$IDF_{t,i}$ 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.

We validate these textual measures using two settings where higher effort is anticipated. First, report length and complexity are higher when evaluators have stronger links to the candidate (e.g., prior co-authorship), even after conditioning on candidate and evaluator fixed effects (Appendix Table A8). Second, reports are longer and more complex for marginal candidates whose qualification hinges on a single vote, even after conditioning on candidate's quantity and impact of publications and evaluator fixed effects, consistent with greater expected scrutiny (Appendix Table A9). While these patterns support interpreting textual characteristics as effort proxies, they capture only one dimension of evaluators' work – time devoted to written assessment – and omit other aspects of evaluation.

4 Empirical analysis

We examine the role of more accomplished experts – researchers with stronger publication records – in shaping evaluation outcomes. First, we analyze how the aggregate expertise of committee members influences evaluation criteria and selection standards. Second, we assess whether committees with greater overall expertise reach better decisions, as reflected in the subsequent research impact and career progression of selected candidates. Third, we examine how expert evaluators influence the decision-making of committees. Specifically, we analyze whether they change the voting behavior of fellow evaluators and whether they affect the effort exerted in evaluation by their peers. Finally, we estimate the costs of committee service, examining how participation affects evaluators' own research productivity and their subsequent willingness to serve.

4.1 Committee expertise and evaluation criteria

We examine whether committees that are randomly composed of better-published researchers apply stricter standards and emphasize different aspects of candidates' research records. In particular, we test whether they attach greater importance to publication impact relative to quantity.

We estimate the following equation on all candidates who pre-registered their applications:

$$\begin{aligned} \mathit{Qualified}_{i,e} = {} & \beta_0 + \beta_1 N_i + \beta_2 Q_i + \beta_3 T_e + \beta_4 T_e \times N_i + \beta_5 T_e \times Q_i \\ & + \beta_6 \mathbb{E}\left[T_e\right] + \beta_7 \mathbb{E}\left[T_e\right] \times N_i + \beta_8 \mathbb{E}\left[T_e\right] \times Q_i + \mathbf{X_{i,e}}\gamma + \varepsilon_{i,e} \end{aligned} \tag{1}$$

where the outcome indicates whether candidate i is qualified in exam e. N_i is the total number of publications in the previous ten years, and Q_i is the share of high-impact publications over the same period. Committee quality, T_e, is defined as the average publication quality of all committee members, T_j, in exam e.

The expected committee quality, $\mathbb{E}[T_e]$, is computed by simulating one million times the official draw procedure, computing each evaluator's probability of being selected $\mathbb{P}\left[drawn_{j,e}\right]$, and then weighting T_j by these probabilities:

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

Candidate-level measures (N_i, Q_i) are normalized within each exam, and evaluator expertise (T_j) within each pool of eligible evaluators. With these normalizations, β_1 and β_2 capture the change in qualification probability associated with a one–standard-deviation increase in candidates' publication quantity and impact, respectively. β_3 measures the effect of being evaluated by a committee whose members are all one standard deviation better published, while β_4 and β_5 capture how committee expertise shifts the marginal returns to quantity and impact.

Identification exploits the random draw of evaluators: conditional on the pool of eligible volunteers, deviations of realized from expected committee composition reflect lottery variation.[28] Because some initially drawn evaluators resigned and were replaced, we instrument final committee quality (and its interactions) with the initial-draw composition and estimate the model by two-stage-least-squares (2SLS).[29] In some specifications, the covariate vector $\mathbf{X_{i,e}}$ includes academic-proximity controls between the candidate and the committee (same subfield, prior coauthorship, same affiliation) to rule out that results reflect pre-existing ties. Standard errors are clustered at the exam level.

Table 1 reports the results. Both publication quantity and impact are positively associated with qualification (column 1): a one–standard-deviation increase in quantity and in the share of high-impact articles raises the probability of qualification by 12 and 9 percentage points, respectively.

More strikingly, we find that committees with greater expertise apply stricter standards. A one–standard-deviation increase in average committee quality reduces the probability of qualification by 6 percentage points. Since committees consist of four Italy-based members, replacing

[28] Randomization checks are reported in Appendix Table D1.

[29] First-stage estimates and F-statistics are reported in Appendix Table D2.

one evaluator with a colleague who is one standard deviation better published lowers an average candidate's success rate by about 1.5 percentage points (4% of the 37% baseline).

Appendix Table A10 reports an alternative specification that replaces average committee quality with the number of *top researchers* serving on the committee. Each additional top researcher is associated with a further reduction in qualification probability, indicating a monotonic relationship between committee expertise and evaluation strictness. While point estimates suggest that marginal effects may decline as additional top researchers are added, the estimates are not precise enough to characterize the functional form. Overall, these results support the use of a continuous measure of committee quality in the baseline analysis.

In column 2 of Table 1, we control for candidate-evaluator connections to ensure that the results are not driven by better-published evaluators being more connected to better candidates. The findings are robust to controlling for coauthorship, shared affiliation, and research proximity.

The interaction terms in column 3 indicate that higher-quality committees shift emphasis from publication quantity to impact. When one committee member is one standard deviation better published, the marginal effect of quantity falls by 5%, while the marginal effect of impact rises by 9%.[30]

Columns 4 and 5 split the sample by promotion level. The increase in strictness and the reduced marginal returns to quantity are driven primarily by associate-level candidates, while the higher marginal returns to impact are present at both levels. Columns 6 and 7 report results separately for STEM+M and SSH fields; point estimates suggest some differences, but none are statistically significant.

Overall, the results suggest that better-published committees apply different criteria, placing greater emphasis on research impact relative to quantity and adopting stricter standards.

4.2 Committee composition and candidates' outcomes

The tendency of expert evaluators to favor candidates with fewer but higher-impact publications can be interpreted in several ways. It may reflect preferences for particular forms of scientific output, such as journal articles over books or other outputs. Alternatively, better-published evaluators may rely more heavily on verifiable signals of research quality, or they may simply be better able

[30] The point estimate on the interaction of publication impact and committee quality is 0.034. This reflects the effect of the entire committee being one standard deviation higher in quality. Dividing by four yields the effect of a single evaluator: 0.85 percentage points. Relative to the baseline marginal effect of publication impact (0.093), this corresponds to an increase of about 9%.

to identify genuinely impactful work. These interpretations lead to a central question: do the change in standards improve the quality of decisions, or do they distort information aggregation by overemphasizing certain signals of research impact?

To address this question, we test whether candidates promoted by better-published evaluators subsequently achieve higher research productivity and stronger career outcomes. Following Zinovyeva and Bagues (2015), we use post-promotion outcomes to capture both observable and unobservable dimensions of candidate potential at the time of evaluation. We estimate the following equation on the sample of qualified candidates:

$$y_{i,e,t} = \beta_0 + \beta_1 T_e + \beta_2 \mathbb{E}\left[T_e\right] + \mathbf{X_i}\gamma + \varepsilon_{i,e,t} \tag{2}$$

where $y_{i,e,t}$ denotes the post-qualification outcome observed in year t for candidate i evaluated by committee e. The coefficient β_1 captures whether candidates qualified by randomly more accomplished committees subsequently achieve better outcomes. We also examine whether this relationship can be explained by pre-exam characteristics, $\mathbf{X_i}$, including affiliation, academic rank, publication quantity and impact, and citations at the time of evaluation. This allows us to assess whether better-published committees weight observable measures more effectively or whether they might be better in identifying talent beyond easily observable indicators.

We estimate Equation (2) by 2SLS, instrumenting final committee quality by the initially drawn committees' quality, and we cluster standard errors at the exam level. We interpret a positive estimate of β_1 as evidence that higher-quality committees systematically select candidates with stronger long-run potential. This interpretation relies on the assumption that committee decisions affect candidates' future outcomes only through selection, not through any direct effect of the committee's identity. In particular, we assume that being qualified by a higher-quality committee does not provide additional information to editors or university-level promotion committees beyond the fact of qualification itself. This concern is mitigated by the relatively high qualification rate (37%), which limits the distinctiveness of any single committee's endorsement.

A further concern is that better-published committees qualify fewer candidates (Section 4.1). Stricter standards may yield a smaller pool of, on average, stronger candidates, but this does not necessarily imply better selection. To isolate the selection-criteria channel from stricter thresholds, we compare a comparable number of top-ranked candidates across committees of different quality. Committees above the median evaluator-quality shock qualify 33% of candidates, compared to 40%

below the median. Restricting the latter group to unanimously qualified candidates yields more comparable success rates (33% vs. 35%). In this restricted sample, a larger β_1 is more likely to reflect more effective selection rather than stricter standards.

We first examine research productivity in the ten years following qualification. Panel A of Table 2 reports estimates for the full sample of qualified candidates, and Panel B for the unanimity-restricted sample. In the full sample, committee quality is unrelated to the total number of future publications (Panel A, column 1). However, replacing one evaluator with a colleague one standard deviation better published is associated with a 1.9% of a standard deviation higher share of high-impact publications (column 3), 1.3% of a standard deviation more citations to both pre-evaluation works (column 5), and post-evaluation works (column 7). In economics, these correspond to approximately a 0.5 percentage-point increase in the share of articles in high-impact outlets, 8 additional citations to pre-2012 works, and 18 additional citations to post-2012 works.[31]

Results for the unanimity-restricted sample are slightly smaller and less precise but qualitatively similar: candidates promoted by better-published committees do not publish more in total, but their work receives greater scientific recognition.

We next assess whether these differences can be explained by observable pre-exam characteristics. Controlling for affiliation, rank, publication quantity and impact, and citations at the time of evaluation fully accounts for the association between committee quality and the post-exam share of high-impact publications. By contrast, citation effects persist in both the full and restricted samples. Conditional on observables, candidates qualified by a committee where one member is exogenously one standard deviation better published increases citations to pre-exam works by roughly 0.7% of a standard deviation and citations to post-exam works by about 1% of a standard deviation.

Beyond research output, Figure 1 examines whether candidates selected by higher-quality committees progress more rapidly in their academic careers. Faster promotion provides an additional proxy for candidate quality, capturing dimensions of potential not fully reflected in publication records. Figure 1 relies on the unanimity-restricted sample; Appendix Figure A1 reports corresponding estimates for the full sample. Each panel presents both unconditional estimates and estimates controlling for pre-exam affiliation, rank, and research productivity.

Panel A focuses on qualified Associate Professor candidates. Replacing one average evaluator with a colleague one standard deviation better published increases the probability of promotion within ten years by about 0.9 percentage points (1.3% relative to a 68% baseline), falling to 0.7

[31] Summary statistics on the productivity of qualified candidates are reported in Appendix Table A6.

percentage points once controls are added. Panel B reports analogous estimates for qualified Full Professor candidates; point estimates remain positive but are not statistically significant even with the inclusion of controls. Panel C returns to Associate Professor candidates and examines promotion to Full Professor. A one–standard-deviation improvement in a single evaluator's quality raises the ten-year promotion rate by 1 percentage point in the unconditional specification (5.9% relative to a 17% baseline), and by about 0.9 percentage point after conditioning on pre-exam observables.

Panel C provides strong support for the selection interpretation of our estimates. Committee quality at the Associate Professor stage predicts eventual promotion to Full Professor more than a decade later, under an independent review process. The temporal and institutional separation between evaluations rules out sustained signaling effects and indicates that better-published committees are particularly effective at identifying candidates with long-term scholarly promise.

The persistence of these effects after conditioning on all pre-exam observables implies that high-quality committees do more than optimally re-weight observable productivity metrics. Rather, they appear better at discerning latent dimensions of candidate potential that predict subsequent scientific impact and accelerated career advancement.[32]

4.3 Expert influence in committees

Committees composed of better-published members may apply different standards and reach higher-quality decisions through several mechanisms. In this subsection, we distinguish between two channels. First, experts may directly influence outcomes through their own voting behavior, applying stricter or otherwise distinct standards. Second, their presence may generate peer effects within the committee, inducing other evaluators to adjust their behavior – either by adopting similar evaluative criteria or by exerting greater effort.

We study these channels by analyzing behavior at the evaluator–candidate level and by examining how peers' votes and report characteristics respond to the presence of better-published colleagues.

[32] Survey evidence from selected fields and universities suggests that reviewers with stronger bibliometric profiles may rely more heavily on quantitative indicators in their assessments (Langfeldt et al., 2021). To examine whether this pattern holds in our setting, Appendix Table A12 analyzes whether individual reports written by better-published evaluators are more likely to explicitly reference bibliometric indicators. If anything, we find the opposite: better-published evaluators are significantly *less* likely to mention metrics such as the h-index, citation counts, A-list journals, Scopus, or Web of Science.

Experts' influence on voting

To assess how the presence of better-published evaluators affects the voting behavior of other committee members, we estimate the following equation:

$$vote_{i,j,e} = \beta_0 + \beta_1 T_j + \beta_2 T_e + \beta_3 \mathbb{P}\left[drawn_j\right] + \beta_4 \mathbb{E}\left[T_e\right] + \mathbf{X}_{i,j}\gamma + \varepsilon_{i,j,e}, \tag{3}$$

where the outcome indicates whether evaluator j, serving on committee e, casts a positive vote for candidate i. The vector $\mathbf{X}_{i,j}$ includes candidate productivity measures and evaluator–candidate connections, such as belonging to the same subfield, prior coauthorship, or affiliation with the same institution. We estimate Equation (3) by 2SLS and cluster standard errors at the exam level.

Identification relies on the fact that, conditional on an evaluator's probability of being drawn, $\mathbb{P}\left[drawn_j\right]$, and on the expected committee composition, $\mathbb{E}[T_e]$, residual variation in peer quality is random. Including expected-composition controls removes predictable correlations between an evaluator's own quality and that of their peers.

To interpret the coefficients in Equation (3), consider a one–standard-deviation increase in the quality of evaluator $j = 1$, holding all other evaluators' quality fixed. Because committee quality T_e is defined as the average quality of the four members, this change increases T_e by 1/4. The effect on evaluator $j = 1$'s own vote is therefore

$$\frac{\partial vote_{i,j=1,e}}{\partial T_{j=1}} = \beta_1 + \frac{\beta_2}{4},$$

while for any other evaluator $j \neq 1$, whose own quality is unchanged, the effect operates only through committee composition:

$$\frac{\partial vote_{i,j\neq 1,e}}{\partial T_{j=1}} = \frac{\beta_2}{4}.$$

Thus, β_1 captures how an evaluator's own standards vary with their quality, relative to peers, while $\beta_2/4$ captures peer effects – how the quality of a single colleague affects other evaluators' voting behavior.

Table 3 reports the results. Column 1 shows that committees composed of better-published evaluators are less likely to cast positive votes, consistent with the committee-level findings.

Column 2 provides strong evidence of peer effects. A one–standard-deviation increase in the quality of a single evaluator lowers the probability that other committee members cast a positive vote by about 2 percentage points. Once this peer effect is accounted for, an evaluator's own quality does not exert an additional influence on voting behavior, consistent with the prevalence of unanimous decisions.

Column 3 examines how the weight placed on candidates' publication quantity and research impact varies with evaluator and committee quality. The results show that higher committee quality significantly increases the return to research impact while reducing the return to publication quantity. In contrast, an evaluator's own quality does not differentially affect how these characteristics are weighted once committee quality is accounted for. This indicates that better-published evaluators influence evaluation criteria primarily through their presence on the committee, inducing all members to adopt a greater emphasis on impact over quantity, rather than by applying different criteria in their own individual assessments.

In column 4, we estimate a specification where we include candidate fixed effects, comparing evaluators who assess the same candidate. The estimate of own-quality effects are precisely centered at zero, indicating no systematic differences within candidates.

Taken together, these results show that although voting is individual, voting behavior is shaped at the committee level. Better-published evaluators influence outcomes primarily through their contribution to overall committee composition rather than through systematically different voting behavior of their own.

Experts' influence on effort

We next examine how peer quality affects the effort exerted by evaluators. Theoretical predictions are ambiguous. The model of Swank and Visser (2023) predicts higher effort when evaluators expect their work to be scrutinized by more accomplished peers, due to concerns about internal reputation. In contrast, models emphasizing external reputation (Visser and Swank, 2007) or strategic information aggregation predict reduced effort or free-riding in the presence of stronger peers.

Although we do not observe the full range of tasks involved in committee service, we focus on individual evaluation reports, which constitute a central and observable component of the evaluation process around which committees are organized (see Section 2.2).

We estimate Equation (3) using report length and textual complexity as outcomes. These measures are normalized at the macro-field by application-category level.[33] This normalization accounts for large differences in reporting styles across fields, while preserving within–macro-field

[33] There are 86 macro fields and two application categories: associate and full professorship.

variation for identification. The interpretation of coefficients parallels the voting analysis, with outcomes now capturing variation in evaluative effort.

Table 4 reports the results. Consistent with the voting analysis, we find strong peer effects. Holding own quality fixed, evaluators assigned to committees with higher-quality peers write longer reports (column 1). A one–standard-deviation increase in the quality of a single peer increases report length by about 7% of a standard deviation, corresponding to roughly 20 additional words per report.

By contrast, we find no evidence that evaluators with higher research quality themselves devote systematically more or less effort to writing reports. As shown by the partial derivative above, we test whether the total own-quality effect ($\beta_1 + \beta_2/4$) differs from zero. We fail to reject this null hypothesis (p-value $= 0.784$), indicating that better-published evaluators write reports of similar length as the average evaluator.[34]

Higher-quality peers also induce evaluators to write more complex reports. A one–standard-deviation increase in peer quality raises the average IDF of words used by about 6% of a standard deviation (column 2), and accordingly also increases total report complexity (column 3).

These findings provide clear evidence of peer effects operating within committees. Evaluators exert greater effort and adjust their evaluation criteria toward those of higher-quality colleagues when serving alongside them, consistent with models emphasizing internal reputation. By contrast, predictions of reduced effort or free-riding in the presence of stronger peers receive little support in this setting.

4.4 Costs of Committee Service and Volunteering

We quantify the costs of committee service by examining whether being assigned to a committee affects evaluators' subsequent research activity. Our empirical strategy exploits the random assignment of evaluators, comparing researchers who are drawn to serve with those who had similar ex ante probabilities of selection but were not drawn.

We proceed in four steps. First, we estimate the average effect of committee service on research output for the full sample of potential evaluators. Second, we examine heterogeneity by researchers' mode of production, focusing on collaboration intensity. Third, we study whether opportunity costs vary systematically with evaluators' baseline research quality. Fourth, we analyze whether serving

[34] Appendix B presents alternative specifications that explicitly separate own-quality effects from peer effects, yielding results consistent with our interpretation.

on a committee affects other academic activities – notably, PhD supervision – and evaluators' willingness to volunteer in future rounds.

Foregone research output

We begin by estimating the average impact of being drawn to serve on an evaluation committee on subsequent research productivity. Using the universe of potential evaluators in the Italian National Scientific Qualification system between 2012 and 2021, we estimate:

$$y_{j,e,t} = \beta_0 + \beta_1 drawn_{j,e} + \beta_2 \mathbb{P}\left[drawn_{j,e}\right] + \mathbf{X_j}\gamma + \mu_e + \varepsilon_{j,e,t} \tag{4}$$

where $y_{j,e,t}$ denotes the research output of researcher j in year t after evaluation e. The indicator $drawn_{j,e}$ equals one if the researcher is selected to serve on the committee. The coefficient β_1 captures the causal effect of committee service on subsequent research productivity. Identification relies on the random assignment of evaluators, conditional on the probability of selection, $\mathbb{P}\left[drawn_{j,e}\right]$.

To increase precision of our estimates, we control for researchers' pre-exam productivity trajectories ($\mathbf{X_j}$), including their number of publications, share of high-impact articles, and a quadratic in academic age. We include exam fixed effects (μ_e) and normalize all productivity measures at the exam level to account for field-specific differences. To address potential bias from endogenous resignations, we instrument actual committee participation with the initial random draw. We estimate Equation (4) by 2SLS and cluster standard errors at the exam level.

Panel A of Figure 2 shows that committee service reduces publication output by about 5% of a standard deviation over the three years following assignment, corresponding to an average loss of roughly 1.5 publications for the average researcher in our sample. This amounts to about 20% of an average evaluator's annual output.[35]

Panel B of Figure 2 reports estimates using cumulative output, set to zero in the year preceding the exam and accumulated thereafter.[36] Committee service generates a persistent decline in productivity that remains statistically significant for up to five years. Although estimates attenuate thereafter, coefficients remain negative even at longer horizons, indicating potentially lasting disruptions to research trajectories rather than short-lived delays.

Our estimates should be interpreted as a lower-bound of the productivity cost of committee

[35] The standard deviation of annual output is roughly 10 publications and the average annual output is 6.5 (Appendix Table A2).

[36] In pre-periods, the outcome is defined symmetrically as cumulative output from a given year to the year before the exam.

service. They identify the causal effect for researchers who both volunteer to be considered and comply with their random assignment. Participation as an evaluator is voluntary, so the eligible pool is likely selected on anticipated costs: researchers who expect particularly high opportunity costs may opt out ex ante. In addition, our instrumental-variables strategy recovers a Local Average Treatment Effect for compliers, i.e. evaluators whose participation status is affected by the random draw. To the extent that researchers facing especially large costs are more likely to decline service after being drawn, the estimated effect may understate the average productivity loss among all eligible researchers.

Heterogeneity by collaboration intensity. The average effect may mask substantial heterogeneity. Researchers working in larger teams may be better able to absorb time shocks through task reallocation and effort adjustment among collaborators, whereas those working in smaller teams may face sharper productivity losses.

Table 5 reports estimates by quartiles of the baseline team size, measured as the average number of coauthors per paper during the ten years prior to the exam (Columns 2-5). Column 1 reports the average effect for the full sample. The negative impact is concentrated among researchers in the bottom quartile: for these researchers, committee service reduces output by 14% of a standard deviation. Effects for researchers above the 25th percentile are smaller and statistically not different from zero. The confidence intervals indicate that, if any, the negative effects are substantially smaller in magnitude.

Motivated by these patterns, we adopt a specification that allows treatment effects to vary continuously with collaboration intensity:

$$y_{j,e,t} = \beta_0 + \beta_1 \frac{drawn_{j,e}}{\bar{\alpha}_{j,e}} + \beta_2 \frac{\mathbb{P}\left[drawn_{j,e}\right]}{\bar{\alpha}_{j,e}} + \beta_3 \frac{1}{\bar{\alpha}_{j,e}} + \mathbf{X}_j \gamma + \mu_e + \varepsilon_{j,e,t} \tag{5}$$

where $\bar{\alpha}_{j,e}$ denotes researcher j's average number of coauthors per paper in the ten years prior to the draw for exam e. This specification preserves identification based on random assignment while allowing the magnitude of the productivity shock to vary smoothly with team size.

Under this formulation, β_1 captures the productivity loss for a researcher who works alone, with effects proportionally attenuated for larger teams. Given an average of 5.93 coauthors per paper, the implied loss for the typical researcher is $\beta_1/5.93$. Appendix Table D3 shows that this scaled specification dominates the binary treatment in a horse-race regression, consistent with adjustment occurring through coauthor responses.[37]

The specification in Equation (5) has several advantages. First, collaboration intensity appears to be a central dimension of heterogeneity in researchers' ability to absorb time shocks and should be accounted for before exploring other margins. Second, modeling this heterogeneity directly improves statistical power. Third, it avoids reliance on arbitrary sample splits.

Figure 3 plots cumulative effects implied by Equation (5). For researchers who exclusively solo-author, committee service reduces output by approximately 19% of a standard deviation in the first three years, with effects remaining statistically significant for up to eight years. However, solo authors constitute less than 5% of the sample. Applying the inverse scaling to the average researcher in our sample, with an average number of 5.93 collaborators per paper, implies a loss of about 3% of a standard deviation, or roughly 1.2 publications. This estimate is smaller and more precisely estimated than the corresponding average effect from Equation (4). In economics, the implied loss is about 1.9 publications for a researcher who exclusively works alone and about 0.6 publications for the average economist, who typically collaborates in teams of three.

Heterogeneity by evaluator quality. We next examine whether opportunity costs vary systematically with researchers' baseline quality. To do so, we interact the scaled treatment indicator with (i) the continuous measure of quality-adjusted publication output and (ii) the indicator variable for *top researcher*.

Table 6 reports how productivity losses from committee service vary with the continuous definition of evaluator quality. Columns 1–2 focus on cumulative effects in the three years following the exam, while Columns 3–4 extend the horizon to eight years. The results indicate that higher-quality researchers incur larger productivity losses, especially in the short-run. An average researcher that exclusively works alone publishes roughly 19% of a standard deviation less due to being drawn for committee service. The estimated impact is about 1.6 times as large for a researcher that is one standard deviation better-published than the average.

To facilitate an easier interpretation of the magnitude of the shock, Panels A and B of Figure 4 plot estimated productivity losses across the distribution of collaborator team size, separately for *top researchers* and other evaluators. At the average team size, committee service reduces output in the short-run by 5.7% of a standard deviation for top researchers, compared with 1.6% for other

[37] Our conclusions in this section are robust to alternative functional-form assumptions, such as the cost of committee service being inversely proportional to the square or the square root of baseline team size.

evaluators – corresponding to losses of about 2.2 and 0.6 publications, respectively.[38]

While heterogeneity by quality is less precisely estimated at longer horizons, the point estimates suggest that the shock has greater persistence among top researchers. Even eight years after being drawn for committee service, we detect a statistically significant decrease in the cumulative research output of top researchers. A top researcher working in a team of average size produces 3.7% of a standard deviation fewer publications, corresponding to a cumulative long-run loss of approximately 1.9 publications.[39]

The largest losses, however, are borne by top researchers working in small teams. Among sole authors, committee service reduces output by as much as 33% of a standard deviation in the three years following the start of service, compared with about a 10% decline for other sole authors. These losses also represent a substantially larger share of top researchers' typical production, reflecting both higher opportunity costs of time and limited scope for adjustment through coauthors.

Overall, these results show that committee service imposes disproportionately large and persistent opportunity costs on better-published researchers, particularly when their mode of production limits the ability to smooth time shocks.

Finally, Panel A. of Appendix Table A13 examines whether productivity losses depend on the quality of randomly assigned peers. We find no statistically significant differences. Although the point estimate is positive, the lower bound of the 95% confidence interval implies that we cannot rule out an additional decline of up to 2.6% of a standard deviation, relative to the average productivity loss of 3% associated with committee participation for the average researcher.

Foregone student supervision

We next examine whether committee service affects PhD supervision. We estimate Equation (5) 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.

[38] In the continuous-quality specification, at the average team size the marginal effect of a one–standard-deviation increase in quality implies an additional short-run loss of about 2.3% of a standard deviation (0.136/5.93), relative to a baseline loss of about 3.3% (0.193/5.93). Given the functional-form assumptions, the statistical significance of the difference does not depend on team size.

[39] Appendix Figure A2 plots the estimates for the accumulated cost over time by different types out outputs.

Unlike research output, we find no meaningful heterogeneity in supervision losses by researcher quality (Appendix Figure A3). Consistently, descriptive statistics in Appendix Tables A1 and A2 show that the number of PhD graduates supervised is broadly similar across researchers of different productivity levels. This pattern suggests that committee service affects the *number* of students supervised in a similar way across researcher types, while leaving open the possibility that adjustment occurs along other dimensions of supervision quality or effectiveness that we do not observe.

Volunteering

Finally, we examine whether the costs of committee service affect evaluators' willingness to volunteer in future rounds. Figure 6 presents descriptive trends in participation rates. Conditional on eligibility,[40] top researchers were initially more likely to volunteer than their less-published peers. Over time, however, their propensity to volunteer declines steadily.

Table 7 provides causal evidence on how prior service affects subsequent participation. We estimate Equation (4) using as the outcome an indicator for whether an evaluator volunteers in the next wave in which participation is ex ante legally permitted for all members of the relevant pool. For evaluators volunteering in 2012, the next eligible wave is 2018; for those volunteering in 2016, it is 2023. Evaluators drawn in 2018 or 2021 are excluded, as they are not eligible to participate in any subsequent observed wave.

A key challenge in this analysis is that participation rules generate a mechanical link between prior service and future eligibility. Evaluators who are not drawn in an early wave remain eligible for intermediate waves, in which they may volunteer and be drawn to serve. If this occurs, they are mechanically ineligible for the later wave used as the outcome. As a result, comparing drawn and non-drawn evaluators in the full sample conflates behavioral responses with a purely rule-based effect.

Column 1 of Table 7 reports estimates from this unrestricted sample. Consistent with the presence of mechanical eligibility effects, the coefficient on prior service is positive but statistically insignificant, and its sign is difficult to interpret.

To isolate the behavioral effect of committee service on subsequent willingness to volunteer, we restrict the sample by excluding evaluators who were not drawn in the initial wave but were drawn to serve in intermediate waves (2016 for the 2012 cohort, and 2018–2021 for the 2016 cohort).

[40] Eligibility criteria evolved across cycles. Appendix Figure A4 reports eligibility rates separately for top researchers (as defined in Section 2.1) and others.

This restriction removes the mechanical eligibility channel by construction. While it may introduce some selection – since evaluators with a higher latent propensity to volunteer are more likely to be drawn in intermediate waves – the resulting bias operates through sample composition and is likely substantially smaller than the mechanical effect it eliminates.

Columns 2 and 3 report results from this restricted sample. We find that prior service significantly reduces subsequent participation, with particularly large effects among highly productive evaluators. Evaluators who were randomly drawn for committee service are roughly 3 percentage points less likely to volunteer in subsequent waves. Among authors with pre-exam productivity one standard deviation above the mean, the reduction in future participation is an additional 3.4 percentage points.

These results point to endogenous selection out of committee service among more productive researchers. Experiencing committee service appears to reduce subsequent willingness to participate, consistent with the substantial opportunity costs documented above.

5 Conclusion

This paper provides causal evidence on how expert evaluators shape collective decision-making and on the opportunity costs that such service entails. Exploiting random assignment of evaluators in Italy's national academic promotion system, the *Abilitazione Scientifica Nazionale*, we show that committees composed of better-published researchers rely on different criteria and reach systematically different decisions. These committees apply stricter standards, place greater weight on research impact relative to quantity, and select candidates who subsequently produce more influential research and advance further in their careers.

We also document the mechanisms through which expertise matters. Better-published evaluators affect outcomes not only through their own votes, but by reshaping committee dynamics. Their presence induces peers to exert greater effort and to converge toward more demanding evaluation criteria, consistent with committee-level reputational pressures. These peer effects imply that the contribution of expertise extends beyond individual judgments to the collective functioning of the committee.

At the same time, expert participation is costly. Serving on a promotion committee causes a persistent decline in evaluators' research output, equivalent to roughly 20% of a typical year's publications. These costs are especially large for highly productive researchers and for those working in smaller teams, who have limited scope to smooth time shocks through collaboration. Committee service also reduces PhD supervision and, crucially, lowers subsequent willingness to volunteer – particularly among the very researchers whose presence most improves evaluation quality. Participation in expert service is therefore endogenous and declines with repeated exposure to its costs.

Our findings highlight a trade-off in the design of evaluation systems. Involving stronger experts improves selection, both directly and through peer effects, but it diverts scarce time away from research and risks adverse selection over time. Systems that rely heavily on expert judgment without addressing its costs may gradually lose precisely the evaluators who contribute most to decision quality.

Our results also point to several implications for institutional design. First, because expert evaluators generate peer effects, their marginal contribution is likely to be highest when expertise is scarce – for example, on committees with few highly accomplished members. This implies that allocating expert time strategically can yield substantial efficiency gains and that evaluation procedures should allow peer effects and reputational incentives to operate. Second, the costs of service are highly uneven: they fall disproportionately on productive researchers working in small teams. Evaluation systems may therefore benefit from limiting repeated service by these researchers or offsetting their higher opportunity costs through compensation, recognition, or service credits. More broadly, expert judgment may be most valuable when deployed selectively, where its marginal value is highest, while using such instruments as transparent indicators, staged review, distributed evaluation, or even partial randomization to reduce evaluation burdens (DORA, 2012; Hicks et al., 2015; Mervis, 2016; Pearson, 2025; VolkswagenStiftung, 2017).

Although our analysis focuses on academic promotion committees, the forces we document are not specific to academia. Many expert-based institutions – from grant panels and journal editorial boards to regulatory agencies and professional licensing bodies – rely on the repeated participation of highly skilled professionals whose time is also highly valuable elsewhere. Our results suggest that, 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.

References

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

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

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

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

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

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

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

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

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

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

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

Dance, A. (2023, February). Stop the peer-review treadmill. I want to get off. *Nature 614*(7948), 581–583.

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

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

Feddersen, T. J. and W. Pesendorfer (1996). The swing voter's curse. *The American economic review*, 408–424.

Funk, P., N. Iriberri, and N. Venus (2024, July). Women in Editorial Boards: An Investigation of Female Representation in Top Economic Journals. *CEPR Discussion Papers* (19303).

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

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols (2015, April). Bibliometrics: The Leiden Manifesto for research metrics. *Nature 520*(7548), 429–431.

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

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

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.

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

Kolarz, P., A. Vingre, A. Vinnik, A. Neto, C. Vergara, C. O. Rodriguez, K. Nielsen, and L. Sutinen (2023, June). Review of Peer Review. Technical report, Technopolis Group.

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

Langfeldt, L., I. Reymert, and D. W. Aksnes (2021, January). The role of metrics in peer assessments. *Research Evaluation 30*(1), 112–126.

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

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

Mervis, J. (2016, August). NSF tries two-step review, drawing praise—and darts. *Science 353*(6299), 528–529.

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

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

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

Pearson, H. (2025, July). How to speed up peer review: Make applicants mark one another. *Nature 643*(8071), 313–314.

Persico, N. (2004, January). Committee Design with Endogenous Information. *The Review of Economic Studies 71*(1), 165–191.

Prat, A. (2005, June). The Wrong Kind of Transparency. *American Economic Review 95*(3), 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, September). Publons' Global State Of Peer Review 2018. Technical report, Publons, London, UK.

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

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

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

Vesper, I. (2018, September). Peer reviewers unmasked: Largest global survey reveals trends. *Nature*.

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

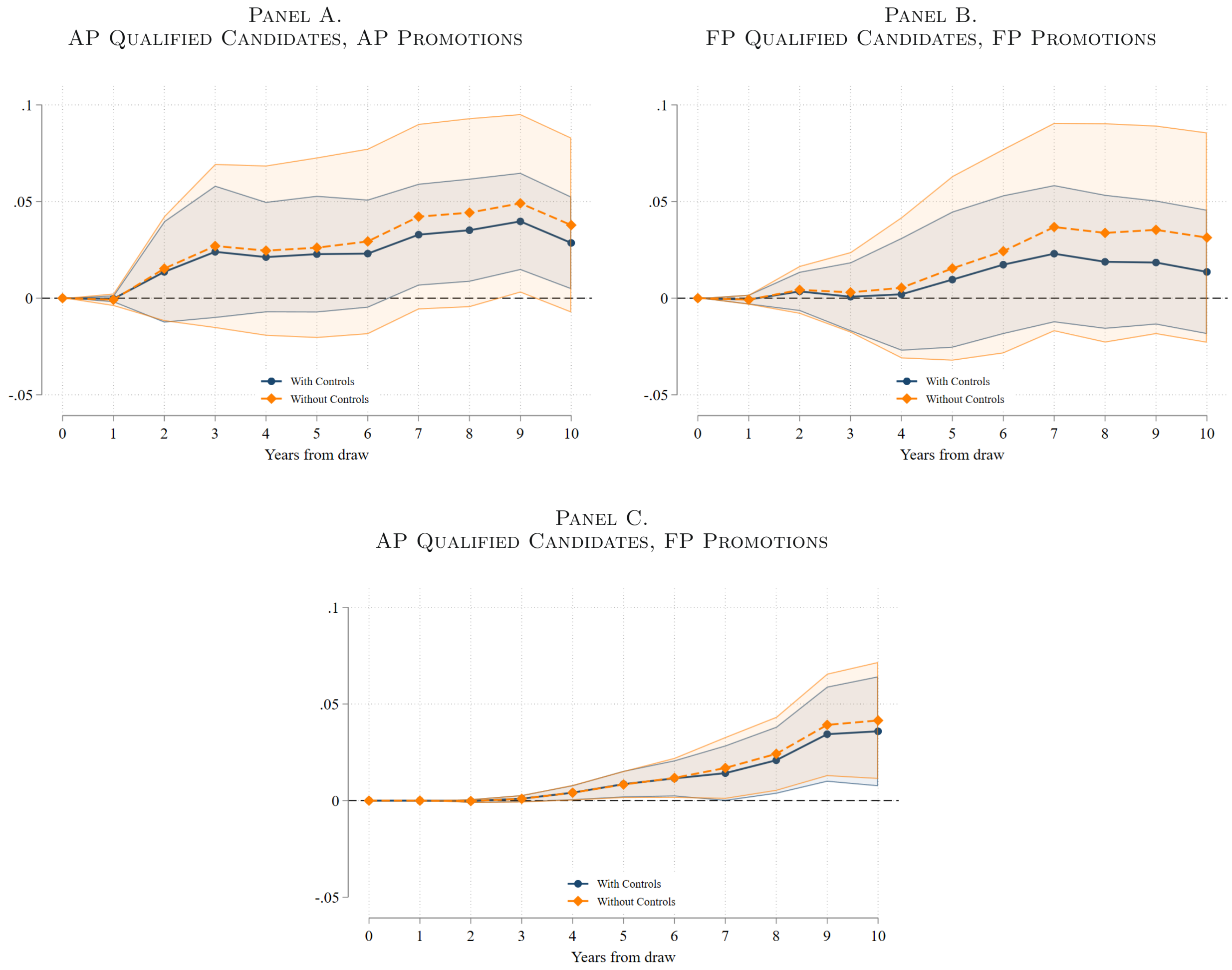
VolkswagenStiftung (2017). Partially Randomized Procedure - Lottery and Peer Review. https://www.volkswagenstiftung.de/en/funding/peer-review/partially-randomized-procedure-lottery-and-peer-review.

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

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

Figures

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



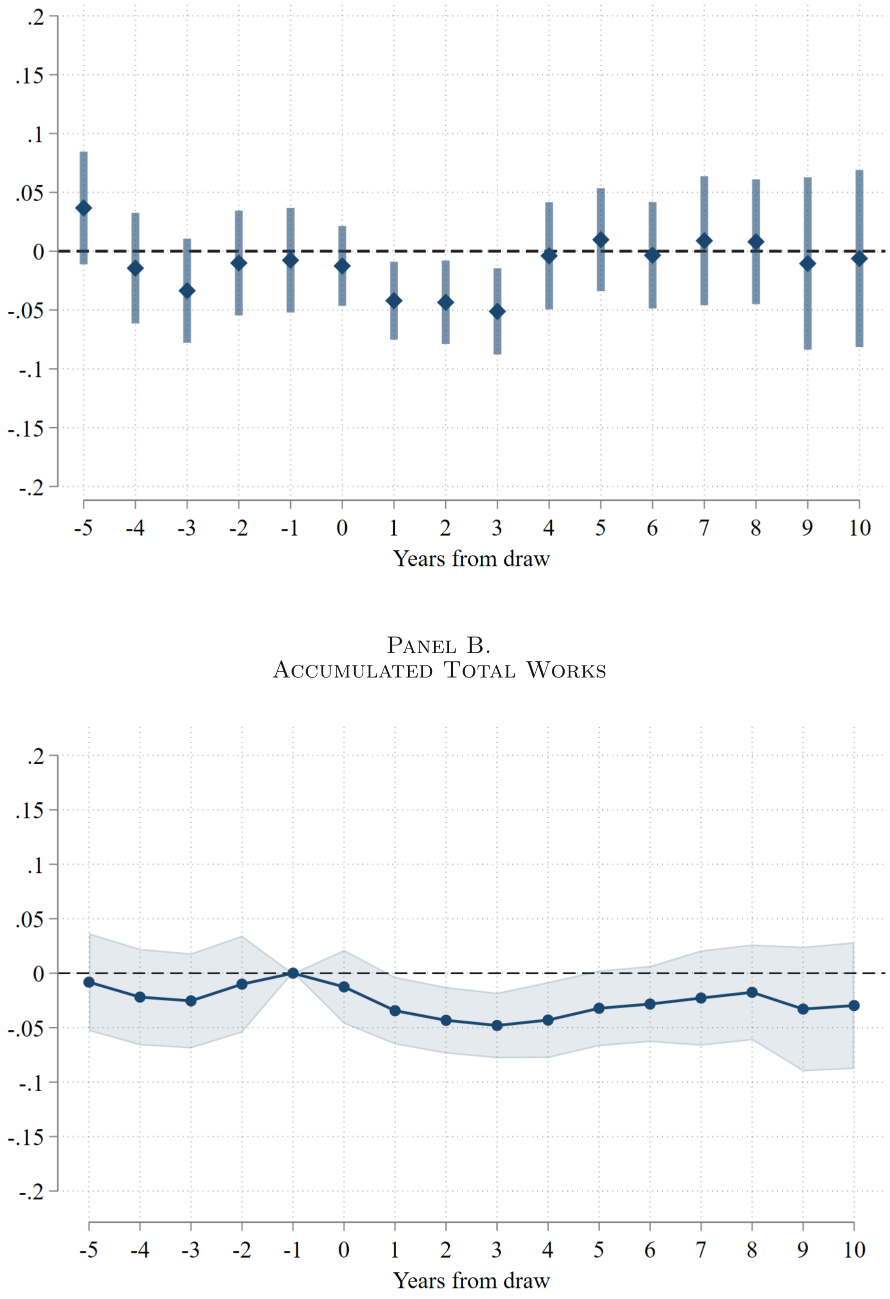
Panel A.
AP Qualified Candidates, AP Promotions

Panel B.
FP Qualified Candidates, FP Promotions

Panel C.
AP Qualified Candidates, FP Promotions

Notes: The sample is qualified candidates in the 2012 wave of the Italian evaluation system. The sample is restricted to unanimous decisions for below-median-quality committees to equalize success rates. Panel A. plots the effect of committee quality on the probability that a top-ranked Associate Professor candidate is promoted to Associate Professor within k years; Panel B. plots the effect of committee quality on the probability that a top-ranked Full Professor candidate is promoted to Full Professor within k years; Panel C. plots the effect of committee quality on the probability that a top-ranked Associate Professor candidate is promoted to Full Professor within k years. Orange dashed lines show estimates without controls; estimates in solid blue lines control for the candidate's pre-exam affiliation, rank, research productivity, and citations. In all figures we control for the committee's expected quality. The final composition of the committee is instrumented by the initial committee composition, due to potentially endogenous resignations. Standard errors are clustered at the exam level. The shaded areas denote the 95% confidence interval around the estimates.

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



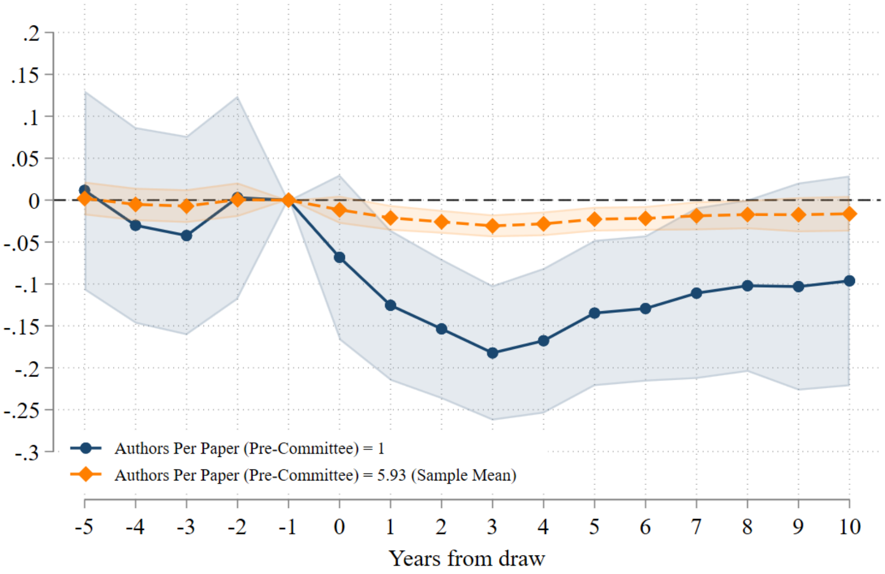
Panel A.
Total Works

Panel B.
Accumulated Total Works

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 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.

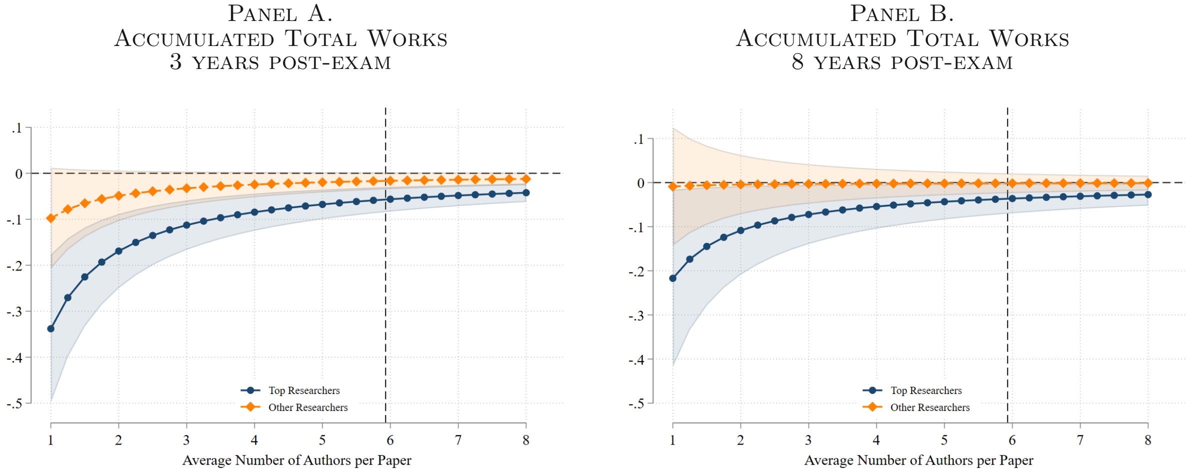
Figure 3: Impact of sitting on the committee on total research output by average number of collaborators per paper prior to the exam

Accumulated Total Works



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 (5). 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



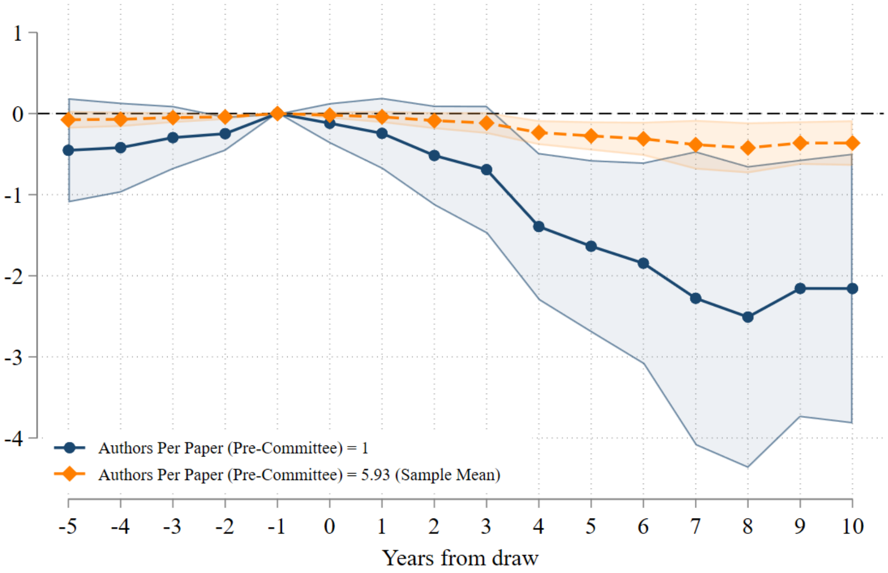
Panel A.
Accumulated Total Works
3 years post-exam

Panel B.
Accumulated Total Works
8 years post-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 (5). 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.

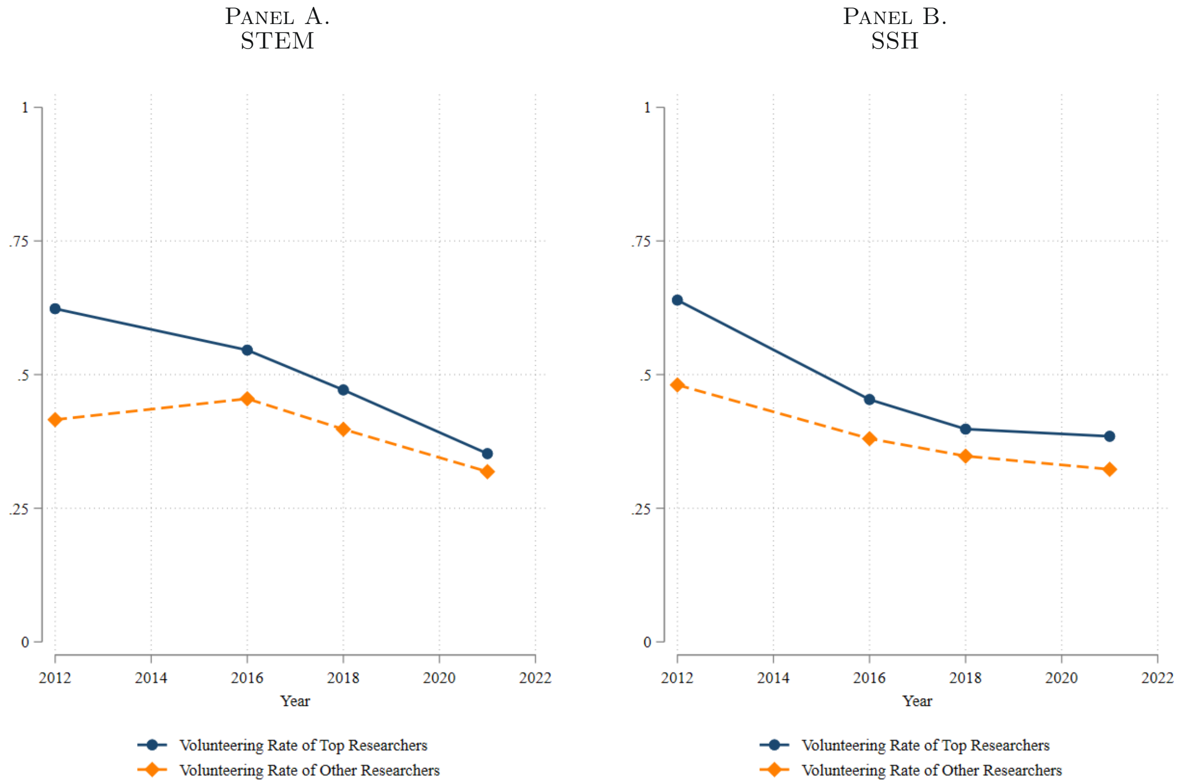
Figure 5: Impact of sitting on the committee on total number of PhD graduates supervised by average number of collaborators per paper prior to the exam

Accumulated PhD Graduates Supervised



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 (5). 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.

Figure 6: Volunteering rate of top researchers



Notes: The sample is all researchers at public universities who were eligible to participate as evaluators in the Italian evaluation system over the period 2012-2021. The variable on the y-axis is the share of eligible researchers who were observed in the pool of potential evaluators. Panel A. reports the volunteering rate in STEM fields. Panel B. reports the volunteering rate in Social Science and Humanities. The solid blue line shows trends for top researchers. The orange dashed 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. We simulate eligibility following the official criteria published by the Ministry.

Tables

Table 1: Committee composition and evaluation outcomes

	Panel A. Overall			Panel B. By Application Level		Panel C. By Field	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				ASP	FP	STEM	SSH
Candidate's Number of Publications	0.121***	0.113***	0.113***	0.120***	0.093***	0.122***	0.099***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.005)
Candidate's Impact of Publications	0.093***	0.093***	0.094***	0.094***	0.092***	0.092***	0.092***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.007)	(0.007)	(0.009)
Committee Quality	-0.061**	-0.065**	-0.065**	-0.073**	-0.047	-0.071**	-0.048
	(0.031)	(0.031)	(0.031)	(0.033)	(0.034)	(0.030)	(0.074)
Candidate's Number of Publications × Committee Quality			-0.023**	-0.029***	-0.008	-0.022	-0.018
			(0.011)	(0.011)	(0.015)	(0.016)	(0.016)
Candidate's Impact of Publications × Committee Quality			0.034***	0.025**	0.055***	0.027*	0.051**
			(0.012)	(0.011)	(0.017)	(0.014)	(0.021)
Observations	69020	69020	69020	47426	21594	41395	27625
Number of Exams	184	184	184	184	184	109	75
Adjusted R^2	0.09	0.14	0.14	0.15	0.13	0.14	0.13
Candidate Connections		✓	✓	✓	✓	✓	✓

Notes: The sample is all candidates who submitted an application in the 2012 wave of the Italian evaluation system. Panel B. splits the sample by application level. Panel C. splits the sample by field. The outcome variable is a binary indicator equal to one if the candidate qualifies. *Candidate's Number of Publications* (N_i) is total research output. *Candidate's Impact of Publications* (Q_i) is the share of candidates' articles in top-quartile Web of Science journals (STEM+M) or A-list journals (SSH). *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. All candidate productivity measures are 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 and – in columns 3–7 – its interactions with candidates' publication quantity and impact. In columns 2–7, we additionally control for the number of connections between evaluators and candidates (coauthors, colleagues, same disciplinary sector). Standard errors are clustered at the exam level.

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

Table 2: Future research productivity of selected candidates

Panel A.
All qualified candidates

	Total Publications		Share High-Impact		Citations (Pre-exam works)		Citations (Post-exam works)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Committee Quality	0.017	0.013	0.076***	0.016	0.053*	0.028**	0.051**	0.040**
	(0.026)	(0.019)	(0.021)	(0.016)	(0.030)	(0.013)	(0.024)	(0.016)
Observations	25342	25337	25342	25337	25342	25337	25342	25337
Adjusted R^2	0.00	0.18	0.00	0.15	0.00	0.61	0.00	0.16
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓
Pre-Exam Observables		✓		✓		✓		✓

Panel B.
Unanimity restricted sample

	Total Publications		Share High-Impact		Citations (Pre-exam works)		Citations (Post-exam works)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Committee Quality	0.004	0.010	0.072***	0.019	0.037	0.024*	0.040	0.037**
	(0.026)	(0.019)	(0.022)	(0.018)	(0.031)	(0.013)	(0.025)	(0.016)
Observations	23776	23772	23776	23772	23776	23772	23776	23772
Adjusted R^2	0.01	0.18	0.00	0.15	0.00	0.61	0.00	0.16
Candidate Connections	✓	✓	✓	✓	✓	✓	✓	✓
Pre-Exam Observables		✓		✓		✓		✓

Notes: Panel A. is estimated on the sample of all qualified candidates in the 2012 wave of the Italian evaluation system. In Panel B. 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. 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. The quality of the committee (T_e) is measured by the average quality-adjusted publications of evaluators over the past 10 years prior to the exam. The final composition of the committee is instrumented by the initial committee composition, due to potentially endogenous resignations. We control for the probability the expected quality of the committee, and connections (coauthors, colleagues, same disciplinary sector). 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 3: Committee composition and individual voting

	Positive Vote			
	(1)	(2)	(3)	(4)
Committee Quality	-0.078**	-0.077**	-0.077**	
	(0.033)	(0.035)	(0.035)	
Evaluator Quality		-0.001	-0.001	-0.000
		(0.005)	(0.005)	(0.002)
Evaluator Quality × Candidate's Number of Publications			-0.000	
			(0.001)	
Evaluators Quality × Candidate's Impact of Publications			-0.002	
			(0.001)	
Committee Quality × Candidate's Number of Publications			-0.023*	
			(0.013)	
Committee Quality × Candidate's Impact of Publications			0.040***	
			(0.012)	
Observations	241744	241744	241744	241744
Adjusted R^2	0.13	0.13	0.13	0.00
Candidate-Evaluator Connections	✓	✓	✓	✓
Candidate Fixed Effects				✓

Notes: The sample is all evaluation reports submitted in the 2012 wave of the Italian evaluation system. The outcome variable is a binary indicator equal to one if the evaluator casts a positive vote. *Evaluator quality* (T_j) is the quality-adjusted publication record of evaluators over the ten years preceding the exam. *Committee quality* (T_e) is 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, due to potentially endogenous resignations. We control for and candidate productivity, and connections (coauthors, colleagues, same disciplinary sector) in all specifications. Additionally, in columns 1-3, we control for the expected quality of the committee; in column 2-4, we control for researchers' (re-centered) probability of being drawn; in column 3, we control for the interaction of expected committee quality and researchers' (re-centered) probability of being drawn with candidates' impact and quantity of publications. Standard errors are clustered at the exam level.

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

Table 4: Committee composition and individual report writing

	Words (1)	Average IDF (2)	Total IDF (3)
Evaluator Quality	-0.081***	-0.088***	-0.090***
	(0.029)	(0.029)	(0.028)
Committee Quality	0.290**	0.260**	0.294**
	(0.129)	(0.114)	(0.127)
Observations	241744	241744	241744
R^2	0.020	0.003	0.016
Candidate-Evaluator Connections	✓	✓	✓
Candidate Controls	✓	✓	✓
$P(\beta_1 + \beta_2 \times 0.25) = 0$	0.784	0.443	0.590

Notes: The sample is all evaluation reports submitted in the 2012 wave of the Italian evaluation system. 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 evaluators over the ten years preceding the exam. *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. $P(0.25 \times T_e + T_j) = 0$ reports the p-value from the Wald test, testing the null-hypothesis that a researcher who is one standard deviation better writes reports of the same length as the average researcher, against a two sided alternative. We control for researchers' (re-centered) probability of being drawn, the expected quality of the committee, candidate productivity, and connections (coauthors, colleagues, same disciplinary sector) in all specifications. The final composition of the committee and the quality of peers is instrumented by the initial committee composition and initial quality of peers, due to potentially endogenous resignations. Standard errors are clustered at the exam level.
*** p < .01, ** p < .05, * p < .1

Table 5: Impact of sitting on the committee on total research output by average collaborator team size

	Accumulated Total Works 3 years post-exam				
	(1)	(2)	(3)	(4)	(5)
	Overall	Q1	Q2	Q3	Q4
Drawn Researcher	-0.048***	-0.140***	-0.034	-0.003	-0.036
	(0.015)	(0.035)	(0.036)	(0.030)	(0.028)
Observations	15691	3687	3788	3949	3889
R^2	0.58	0.50	0.52	0.60	0.70
Average Number of Authors per Paper	5.93	1.71	3.34	5.68	12.62

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 in the first three years following in 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 6: 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 (5). 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 7: Probability of volunteering again

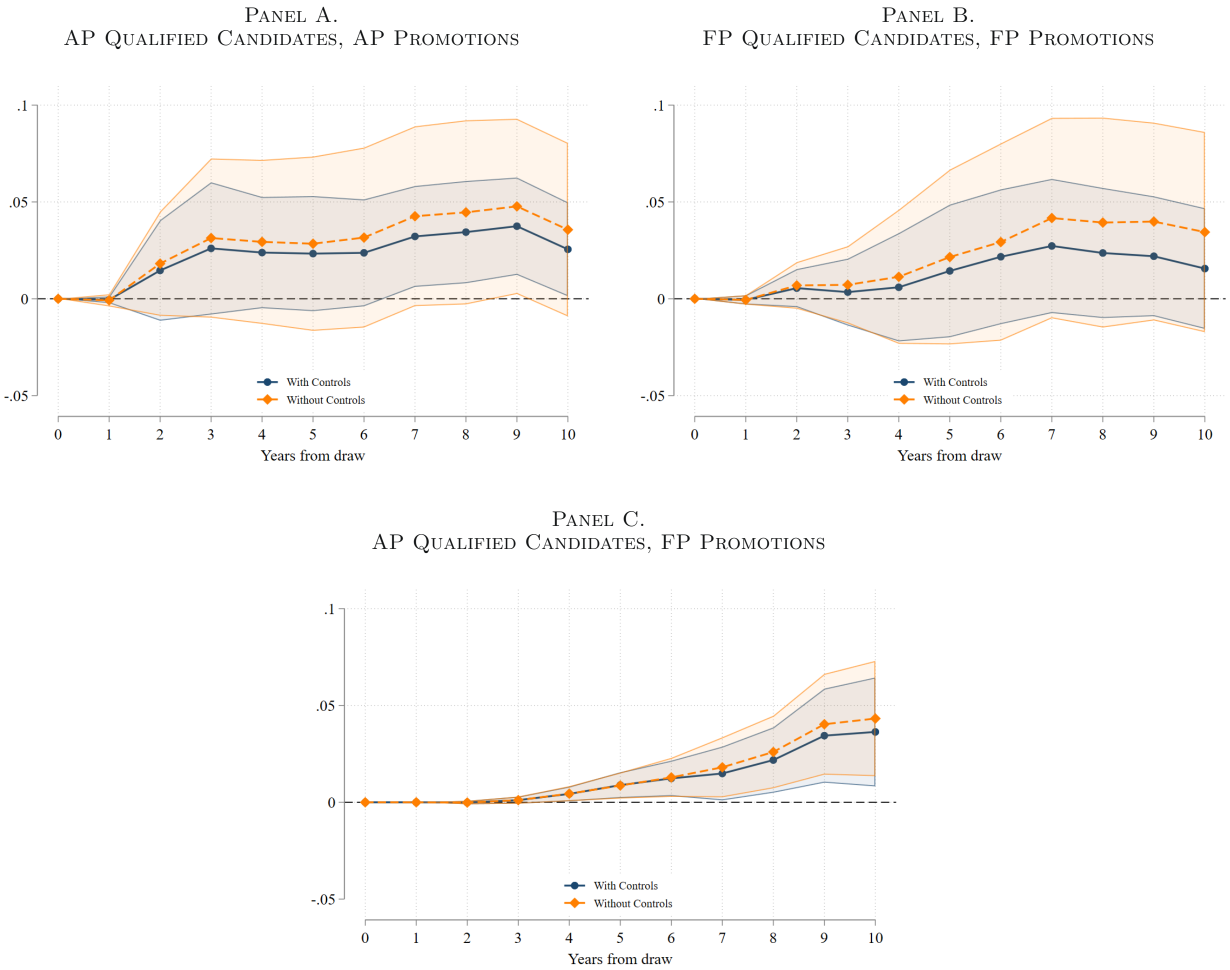
	Volunteer Again		
	(1)	(2)	(3)
Drawn Researcher	0.018	-0.031**	-0.032**
	(0.014)	(0.014)	(0.014)
Evaluator Quality	0.019***	0.023***	0.027***
	(0.004)	(0.005)	(0.006)
Drawn Researcher × Evaluator Quality			-0.034**
			(0.015)
Observations	9552	8696	8696
Adjusted R^2	0.002	0.003	0.003
Restricted Sample		✓	✓
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 2-5, 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 waves 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 wave in which they are eligible to participate (the 2018 wave for 2012 volunteers and the 2023 wave 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

Appendix

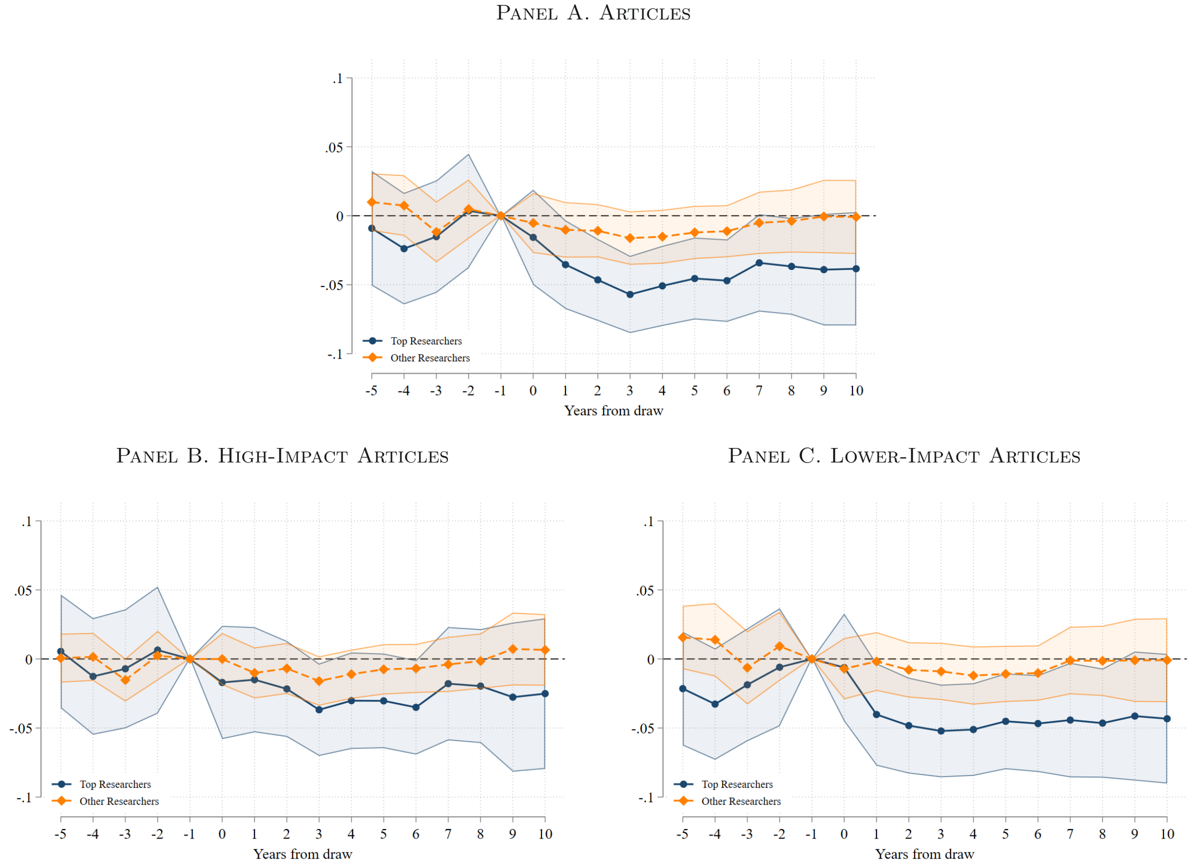
Appendix Figures

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



Notes: The sample is all qualified candidates in the 2012 wave of the Italian evaluation system. Panel A. plots the effect of committee quality on the probability that a top-ranked Associate Professor candidate is promoted to Associate Professor within k years; Panel B. plots the effect of committee quality on the probability that a top-ranked Full Professor candidate is promoted to Full Professor within k years; Panel C. plots the effect of committee quality on the probability that a top-ranked Associate Professor candidate is promoted to Full Professor within k years. Orange dashed lines show estimates without controls; estimates in solid blue lines control for the candidate's pre-exam affiliation, rank, research productivity, and citations. In all figures we control for the committee's expected quality. The final composition of the committee is instrumented by the initial committee composition, due to potentially endogenous resignations. Standard errors are clustered at the exam level. The shaded areas denote the 95% confidence interval around the estimates.

Figure A2: Impact of sitting on the committee on accumulated article production by evaluators' baseline productivity



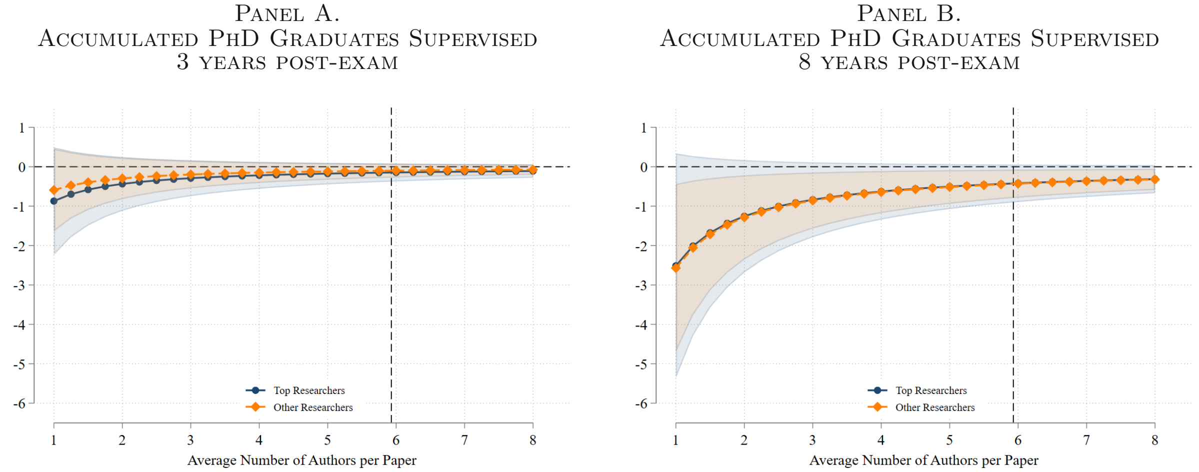
Panel A. Articles

Panel B. High-Impact Articles

Panel C. Lower-Impact Articles

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 (5). 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

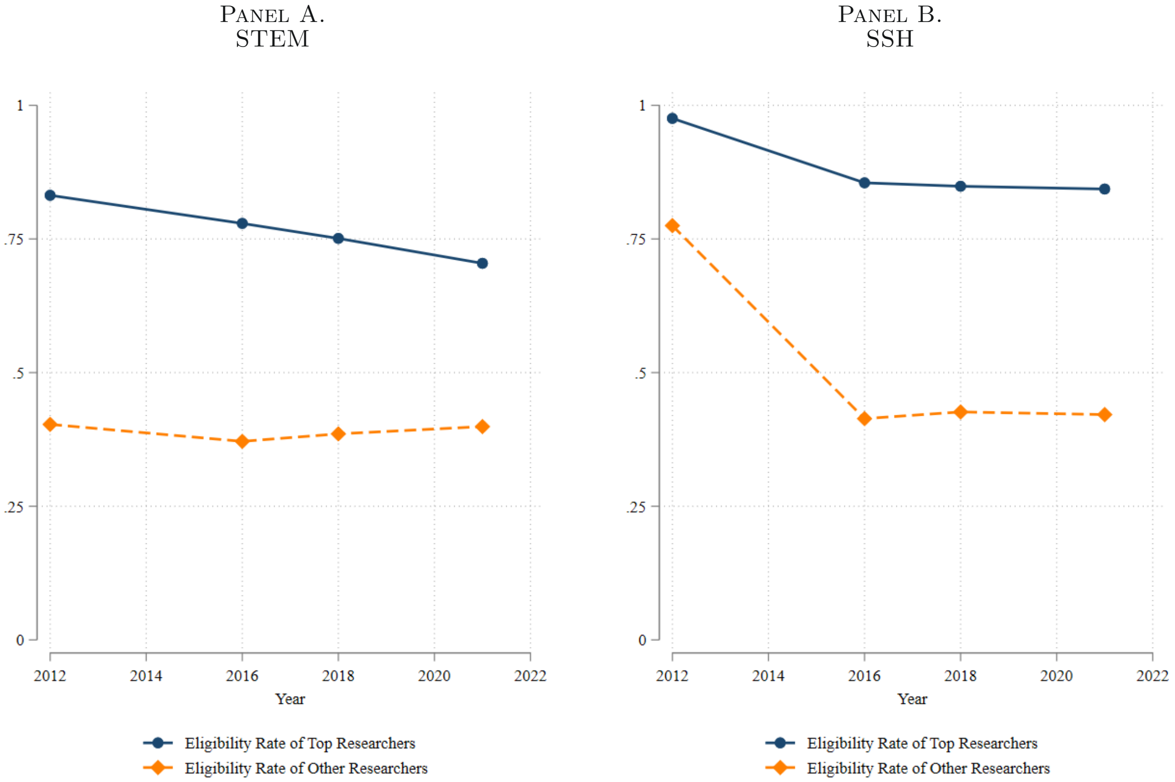


Panel A.
Accumulated PhD Graduates Supervised
3 years post-exam

Panel B.
Accumulated PhD Graduates Supervised
8 years post-exam

Notes: The sample is all researchers who participate in the Italian evaluation system between 2012-2021. The estimates are based on Equation (5). 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.

Figure A4: Eligibility of top researchers



Panel A.
STEM

Panel B.
SSH

Notes: The sample is full professors at public universities over the period 2012-2021. The variable on the y-axis is the share of researchers who were eligible to participate as evaluators in the Italian evaluation system. Panel A. reports the eligibility rate in STEM fields. Panel B. reports the eligibility rate in Social Science and Humanities. The solid blue line shows trends for top researchers. The orange dashed 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. We simulate eligibility following the official criteria published by the Ministry.

Appendix Tables

Table A1: Research productivity in the pool of potential evaluators prior to the exam

	Panel A. Overall		Panel B. Top Researchers		Panel C. 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 A2: Research productivity in the pool of potential evaluators following the exam

	Panel A. Overall		Panel B. Top Researchers		Panel C. 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 A3: Balance tables

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. We simulate eligibility following the official criteria published by the Ministry.
*** p < .01, ** p < .05, * p < .1

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 2012 wave 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

	Panel A. Overall		Panel B. By Field				Panel C. 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 2012 wave of evaluations. The productivity measures are constructed from candidates' CVs. *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 2012 wave 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

	Mean	SD
Positive Vote	0.45	0.50
Words	182.45	285.88
Average IDF	2.54	0.76
Total IDF	304.87	619.88
Observations	241744	
Number of Evaluators	758	
Number of Candidates	40217	

Notes: The sample is all evaluation reports in the 2012 wave of evaluations.

Table A8: Report characteristics and candidate profiles

	Words			Average IDF			Total IDF		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coauthors	0.126**	0.095***	0.115***	0.143***	0.111***	0.117***	0.153***	0.117***	0.134***
	(0.057)	(0.028)	(0.020)	(0.040)	(0.031)	(0.027)	(0.052)	(0.028)	(0.021)
Same University	0.109***	0.090**	0.089***	0.049	0.068**	0.072***	0.105***	0.094**	0.100***
	(0.039)	(0.037)	(0.019)	(0.036)	(0.034)	(0.024)	(0.038)	(0.037)	(0.019)
Same Subfield	0.175***	0.114***	0.128***	0.068**	0.124***	0.093***	0.159***	0.124***	0.138***
	(0.035)	(0.039)	(0.025)	(0.031)	(0.037)	(0.023)	(0.034)	(0.038)	(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)		
Observations	241744	241744	241744	241744	241744	241744	241744	241744	241744
Adjusted R^2	0.013	-0.197	-0.192	0.003	-0.197	-0.196	0.012	-0.197	-0.192
Candidate Fixed Effect		✓	✓		✓	✓		✓	✓
Evaluator Fixed Effect			✓			✓			✓

Notes: The sample is all evaluation reports in the 2012 wave 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 2012 wave 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: 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 2012 wave 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 A11: Non-linear effect of top researchers on future outcomes of top-ranked 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 Researcher on Committee	0.006	0.010	-0.005	0.017	0.059	-0.015	0.060**	0.016	0.005	0.030
	(0.039)	(0.032)	(0.043)	(0.035)	(0.048)	(0.022)	(0.028)	(0.028)	(0.024)	(0.024)
Two Top Researchers on Committee	-0.004	-0.007	0.027	0.033	0.045	-0.007	0.046	0.013	0.036	0.046*
	(0.036)	(0.032)	(0.042)	(0.037)	(0.049)	(0.024)	(0.029)	(0.027)	(0.025)	(0.026)
Three or More Top Researchers on Committee	0.030	0.016	0.053	0.026	0.089*	0.011	0.074**	0.031	0.057**	0.063**
	(0.041)	(0.035)	(0.041)	(0.036)	(0.050)	(0.023)	(0.030)	(0.027)	(0.024)	(0.026)
Obs.	23776	23772	23776	23772	23776	23772	23776	23772	16511	16508
Adjusted R^2	0.000	0.204	0.002	0.172	0.004	0.616	0.003	0.170	0.003	0.116
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 A12: 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 2012 wave 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: 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 waves 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 waves. 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

A Data

We construct our main database by merging information from four sources: (i) the universe of professors at public Italian universities from CINECA, (ii) the pool of potential evaluators and the draw procedures of the national abilitazione system, (iii) the research output of researchers from IRIS and OpenAlex, (iv) data on candidates collected by Bagues et al. (2017).

A.1 Identifying researchers

CINECA provides the core of our database, as it provides complete coverage of all professors employed in Italian public universities. The dataset includes each professor's full name, gender, university, department, field of research, and academic rank. We begin by assigning an identifier to researchers in CINECA, then later we merge the different sources to this database.

Disambiguating professors in CINECA We begin by assigning unique identifiers to each researcher in CINECA to resolve cases of name duplication. We scrape the list of all professors between 2010-2023.

Each observation in CINECA corresponds to a professor-year record. Since the dataset spans multiple years and includes all professors employed in public universities, we assign a unique identifier to each individual to ensure consistent tracking over time.

Most professors have unique full names: in 98% of cases, the full name alone suffices to identify a single individual within a given year. For these observations, we assign an identifier based solely on the full name. For the small number of duplicate names (roughly 2% of observations), we proceed in successive steps.

First, we also consider the research area (*settore scientifico-disciplinare*) to the full name and assign a new identifier. This resolves nearly all remaining cases of homonymy, with only seven instances requiring additional disambiguation. Second, for the remaining seven cases we incorporate the university affiliation into the matching (full name, SSD, and university). We manually validate the accuracy of the matching for these cases. Finally, we assign residual identifiers to any remaining unmatched records based on full name only, however in practice there were no such cases.

We create a single identifier, `cineca_id`, which uniquely identifies professors across years. This procedure ensures that each researcher is uniquely and consistently tracked throughout the database.

Merging ASN databases to CINECA

Linking ASN evaluators to CINECA In the next step, we merge the pool of potential evaluators and the draw status of researchers participating in the national qualification system (ASN) with the CINECA database. Because the population of potential evaluators is drawn from the universe of Italian professors within a given macro field and names are rarely duplicated within a macro field (e.g., economics), homonymy poses limited concern. Therefore, we perform the matching primarily on the basis of names within macro fields, using a sequence of harmonization steps to account for minor inconsistencies in the source data.

The ASN information was extracted from PDFs published by the Ministry. Some PDFs are not supplied in a readable format and need to be processed with OCR. This OCR procedure occasionally introduces errors and capitalization inconsistencies. Before merging, we normalize all names by removing spaces and converting to uppercase. We then generate multiple variants of the name string to account for different possible name orderings (surname followed by first name, first name followed by surname, and combinations with one or two given names).

To correct common scanning or OCR errors, we systematically replace letters that are frequently misread, such as "I" misinterpreted as "T" or "L", and "Q" misinterpreted as "G". We also construct auxiliary keys based on partial matches – for instance, using the surname alone, the surname combined with an initial, or reduced forms of compound surnames – and iterate the merge procedure across these alternatives. After each iteration, we drop duplicate matches to avoid multiple assignments of the same CINECA identifier.

Finally, for a small number of evaluators whose affiliations may have changed between the year of the ASN session and the CINECA reference year, we allow for matching on the previous year's records. This step captures professors who retired, moved universities, or were newly appointed during the evaluation cycle.

The result of the procedure is a correspondence linking virtually all researchers in the ASN pool of potential evaluators to a `cineca_id` (we fail to match 8 researchers from the pool).

Linking ASN candidates to CINECA We next link candidates who applied for qualification in 2012 to the CINECA database. The procedure closely mirrors the one used for evaluators, relying primarily on matching name strings after standardizing capitalization, spacing, and common typographical inconsistencies. However, we adopt a more conservative approach to avoid false positives: we exclude "risky" merge types such as matches based solely on surnames or surname-initial combinations, and we require at least partial match on both given and family names. Because ASN applications are for associate and full professorships, we exclude full professors from the pool of potential matches in CINECA. This restriction ensures that candidates are only matched to individuals who are eligible for promotion.

Using this procedure, we successfully link 33,446 out of 46,329 candidates (approximately 72%) to unique CINECA identifiers. The remaining candidates are mostly based outside the Italian university system, as foreign researchers are not covered by CINECA and therefore cannot be matched.

Assigning research identifiers We construct a researcher identifier that is invariant across sources and years and that accommodates individuals who appear in some but not all datasets. We assign unique identifiers to researchers across three databases: (i) CINECA, (ii) the population of evaluators in ASN, and (iii) the population of candidates in ASN. Nearly all evaluators were matched to CINECA, whereas a substantial share of candidates were not, typically because they apply from outside the Italian university system. Hence, we create an identifier that covers all three data sources.

1. **Appending sources.** We append CINECA, evaluator, and candidate records into a single file, retaining name fields, gender, affiliation, research area, and the source-specific identifiers `cineca_id`, `eval_id`, and `candidate_id`.

2. **Priority key and grouping.** For each record we construct a provisional string key `id_` using the following priority order: `id_` = "cin"+`cineca_id` if available; otherwise "ev"+`eval_id`; otherwise "cand"+`candidate_id`. This prioritizes the CINECA registry when we were able to assign a `cineca_id`, then the evaluator registry, then the candidate registry. We collapse the provisional keys to a numeric `id` via a stable grouping function so that all records sharing the same provisional key receive the same `id`.

3. **Harmonization within `id`.** Within each `id`, we backfill missing source identifiers by carrying nonmissing values across rows: if `cineca_id` is missing for some observations but present for others within the same `id`, we impute the missing entries with the observed value, and analogously for `eval_id` and `candidate_id`. We also standardize incomplete name fields by parsing `fullname` into `surname` and `name` when these are missing.

4. **Researcher roster for merging.** We keep the most informative record per `id` by ranking observations on completeness of affiliation, gender, and area, and we retain the top ranked row. The resulting roster `our_researchers` contains one observation per `id` with standardized name fields, gender, affiliation, and discipline categories. These discipline categories are merged from a correspondence table that maps areas to classifications with different levels of granularity.

This procedure yields a master identifier, `id`, that uniquely tracks individuals across CINECA, evaluator, and candidate records, while preserving links to the original source identifiers. It ensures that researchers who are not present in CINECA, such as candidates applying from abroad, are still consistently represented in the researcher database.

A.2 Merging publication databases to our researchers

We matched the final list of our researchers, where we already assigned unique identifiers across CINECA and the ASN databases to two sources of publication data: IRIS and OpenAlex. We start with IRIS, since it is maintained by the Italian ministry and only contains self-reported publications by Italian researchers. Then, we process OpenAlex following a more complex procedure.

Linking researchers to IRIS

We link the complete list of researchers to publication records from the Italian institutional repository system (IRIS), where professors affiliated with Italian universities self-report their research output. We scrape IRIS twice: once in 2020 and again in 2024. We combine the two datasets to maximize coverage. Before merging, we make sure that there are no duplicates both within and across scrapes, ensuring that each publication corresponding to a specific author profile appears only once.

The merge proceeds in several stages. We first standardize author names by removing spaces, converting to uppercase, and correcting common typographical inconsistencies. For each IRIS author profile, we keep only those with a valid university identifier and one observation per individual-university pair. We then link IRIS authors to our database of researchers using the same logic employed for the CINECA–ASN merges, matching on cleaned names and university affiliations.

To ensure match quality, we classify potential matches according to the uniqueness of names within CINECA. When a name is unique nationwide, we merge directly on the full name. When a name is unique only within a university, we merge on the combination of name and university. For non-unique names, we rely on additional information such as multiple affiliations recorded in

CINECA or IRIS, taking as correct the match whose affiliations coincide. We also assign priority to matches obtained through the most reliable merges (for example, exact string).

After these steps, we discard remaining ambiguous cases and profiles with very few publications (three or fewer), which typically correspond to coauthor entries. The resulting dataset links researchers in our database to a unique IRIS author profile. This provides detailed publication histories for evaluators and candidates, which we later complement with information from OpenAlex to address potential issues with self-reporting.

Add descriptives

Linking researchers to OpenAlex

We take a multi-stage approach to merging our researchers with the OpenAlex database. The procedure begins with a broad set of name-based matches and then applies successive filters to remove implausible links and confirm correct matches.

Name-based initial merge We begin the linkage by constructing a set of potential author matches for each researcher in our master database. This first stage relies solely on names, applied after string normalization.

Name cleaning. We clean all person names to a common format: we convert to uppercase, remove diacritics and punctuation, standardize whitespace and hyphens, and harmonize surname particles (e.g., *De*, *Di*, *Della*, *Del*, *Van*, *Von*). We create ordered variants to accommodate both ⟨given name, surname⟩ and ⟨surname, given name⟩ conventions, and generate initials for the first and second given names when present.

Matching procedure. We then execute a sequence of merge routines, ordered from the most to the least restrictive:

1. **Exact matches:** merge on fully standardized name strings.
2. **Minor variations:** allow for differences in middle names, initials, or multiple given names by dropping or reordering components.
3. **Surname and particle harmonization:** handle variations in surname particles and compound surnames.
4. **Reversed order:** account for cases where surname and given name order is inverted.

Each merge generates a set of potential matches, which are stored in temporary files for subsequent validation. When multiple OpenAlex author records correspond to the same name, we retain the longest name variant to ensure consistency in later matching steps.

In subsequent stages, we discard matches that display conflicting disciplinary information and confirm remaining matches using observable characteristics such as affiliation and research area.

Discarding implausible merges We first refine the list of potential matches obtained from the name-based routines. All temporary merge outputs are appended into a single dataset. For each pair of researcher and OpenAlex author identifiers, we retain the most plausible match, prioritizing stricter merge types and longer name strings.

To eliminate cases that cannot be reliably disambiguated, we discard pairs involving individuals linked to an excessively large number of profiles. Specifically, we remove cases where a researcher (or an OpenAlex author) is matched to more than one hundred distinct counterparts, unless the link originated from an exact string match.

We then remove pairs that are unlikely to be valid matches based on other parts of the full name that may not be utilized by some merge function. We drop OpenAlex records with extremely long person strings and flag conflicts that arise when name initials are inconsistent across sources. To avoid discarding true matches due to minor typos, we compute string distance between our cleaned name and the OpenAlex name and retain pairs with very small differences.

Confirming correct merges In this stage, we augment profiles with additional information and confirm correct merges.

Enriching with observables. For each candidate pair we merge in a set of characteristics used for disambiguation and confirmation:

- affiliation: we search for the researcher's home institution within the OpenAlex affiliation history, including synonyms and components extracted from the registrar source;
- journals: we check whether journals listed by the researcher appear in the OpenAlex publication list;
- country: we extract countries from the OpenAlex affiliation history;
- discipline: we construct both a narrow discipline code and a broader four–category field on both sides and compare them.

Confirming matches. We then confirm links in a sequence that proceeds from most to least demanding criteria. Each rule uses a uniqueness requirement: a pair is confirmed only when exactly one OpenAlex profile and exactly one researcher satisfy the rule simultaneously. Once a pair is confirmed, we remove that OpenAlex profile from consideration for all other researchers. The sequence is:

1. **Unique one to one by name.** If a researcher is linked to only one OpenAlex profile and that profile links to only that researcher, we confirm the pair.

2. **Affiliation plus strict name variant.** Among pairs that also satisfy a strict name variant, we confirm when the affiliation matches and the one to one condition holds within this subset.

3. **Affiliation only.** We confirm when the affiliation matches and the one to one condition holds within the affiliation–matched subset.

4. **Journal evidence.** We confirm when there is at least one journal overlap between the CVs and OpenAlex, and the one to one condition holds within the journal–matched subset.

5. **Strict name plus discipline.** We confirm when the strict name variant and the narrow discipline coincide and the one to one condition holds within this subset.

6. **Strict name plus discipline and country.** We confirm when the strict name variant, discipline, and country all coincide and the one to one condition holds within this subset.

7. **Strict name plus broad field and country.** We confirm when the strict name variant, the broader four–category field, and country coincide and the one to one condition holds within this subset.

8. **Country with auxiliary field information.** As a final pass for remaining cases, we confirm when country coincides and the one to one condition holds within subsets defined by narrow field, broad field, or strict name variant, respectively.

Conflict resolution. At each step we ensure that a confirmed OpenAlex profile is not assigned to multiple researchers by removing it from all other candidate sets. The procedure alternates between confirmation and pruning so that later, weaker rules operate only on unresolved cases. Throughout, we avoid relying on a single signal. Confirmation requires agreement between names and at least one independent piece of evidence such as affiliation, journals, discipline, or country, together with uniqueness within the relevant subset.

Output. The result of this stage is a set of confirmed links that satisfy deterministic criteria: a researcher and an OpenAlex profile are linked only when the pair is uniquely supported by observables drawn from both sources.

Add descriptives

Combining IRIS and OpenAlex publications

We construct a unified publication database by merging data from IRIS and OpenAlex, harmonizing variable definitions, and removing duplicates across the two sources. This database forms the basis for all productivity measures used in the analysis.

Harmonization and merge. We begin with the IRIS publication database and harmonize variable names and publication-type codes to match OpenAlex conventions. We then append OpenAlex publications, retaining only substantive work types such as journal articles, books, book chapters, and conference proceedings.

Title cleaning. We standardize publication titles by removing punctuation, line breaks, and diacritics, truncating them to one hundred characters, and applying a Python-based cleaning script that removes stopwords and normalizes spacing. For efficient duplicate detection, we retain a compressed version of each title containing the first ten words.

De-duplication across sources. The IRIS and OpenAlex databases contain overlapping records for many publications, reflecting both common coverage and differences in indexing practices. We therefore implement a multi-step de-duplication procedure that identifies and removes duplicate entries while preserving the most informative record for each work.

First, we standardize titles across the two sources and also harmonize author lists, publication years, and identifiers such as ISSN and DOI where available.

Second, we identify potential duplicates using a hierarchical matching strategy:

1. **Exact title and year matches.** We flag pairs of IRIS and OpenAlex records that share an identical cleaned title and publication year.

2. **Fuzzy title matches.** For titles differing only by minor details (for example, missing accents, hyphens, or reordered words), we compute string distances and retain matches below a small tolerance threshold.

3. **Metadata confirmation.** Among candidate pairs, we confirm duplicates when at least one of the following fields coincides: DOI, ISSN, or journal name. We also check for overlapping authorship, defined by sharing at least one author surname–initial combination.

Third, we remove redundant observations based on a source hierarchy. When a publication is present in both databases, we keep the OpenAlex entry, which typically contains richer affiliation and author identifiers, and drop the corresponding IRIS record.

Finally, we verify that each unique title–year–author combination appears only once in the unified database. The resulting dataset contains a single observation per publication, with metadata harmonized across sources. This ensures that productivity measures reflect unique research outputs rather than multiple database representations of the same work.

Add descriptives

A.3 Simulating the ASN draw procedure

2012 ASN The draw of evaluators for the 2012 ASN followed a protocol designed to ensure representation across research areas and institutions. Each draw relied on two pieces of information about potential evaluators: their affiliation and their research discipline (*settore scientifico-disciplinare*). Information on affiliation is directly available in the official draw files, whereas SSDs are merged from CINECA. For eight evaluators with missing SSD, we impute this information using the mode SSD within the corresponding field.

The official draw procedure can be summarized in three main steps:

1. **Ordering by SSD size.** SSDs are first sorted by the number of volunteers in ascending order, so that smaller SSDs are considered before larger ones. This ordering ensures that evaluators from smaller fields are not systematically excluded due to institutional constraints that arise later in the process.

2. **Ensuring representation of large SSDs.** From each large SSD (more than 30 researchers in CINECA) in the ordered list, one evaluator is drawn sequentially. Within each SSD, evaluators are examined in random order, and the algorithm selects the first candidate whose institution has not yet been drawn. Once a candidate is selected, both their SSD and institution are marked as represented.

3. **Completing the draw.** After one evaluator has been drawn from each SSD, any remaining slots are filled from the overall pool of eligible evaluators. This final draw proceeds in random

order, again ensuring that no two evaluators from the same university are selected. The process continues until the required number of evaluators is reached.

We simulate this process one million times to recover each researcher's ex ante draw probability. We also store the probability that any two researchers get drawn on the same committee. This probability will become important when we compute peer effects, since it will allow us to control for the expected peer quality.

2016-2023 ASN The draw procedure for the subsequent ASN rounds (2016-2023) followed the same principles as in 2012 but introduced several modifications aimed at improving the proportionality of representation across fields. Two main differences distinguish this period. First, the threshold defining a "large" SSD was reduced from 30 to 10 eligible full professors, expanding the set of SSDs guaranteed representation in the initial stages of the draw. Second, the selection from large SSDs became proportional to field size, rather than assigning a single slot to each SSD irrespective of its number of volunteers.

Formally, SSDs with at least 10 eligible evaluators are classified as large and ordered by size in descending order. For each SSD, we compute the expected number of evaluators as

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

where N denoted the total number of evaluators drawn for an exam, S denotes the set of SSDs within a discipline, and N_s denotes the number of eligible professors in SSD s. E_s is the expected number of evaluators drawn from the SSD. The number of evaluators who are guaranteed to be drawn from the SSD is equal to $floor(E_s)$.

The algorithm proceeds in three stages.

1. **Iteration 1: Ensuring representation of large SSDs.** Evaluators are drawn from each large SSD (more than 10 researchers in CINECA) in a first round, giving priority to the largest fields and assigning one evaluator whenever the expected number E_s exceeds one. As in 2012, selection within each SSD respects the institutional constraint that no two evaluators come from the same university.

2. **Iteration 2: Ensuring representation of large SSDs.** The algorithm iterates over large SSDs to fill remaining expected slots until the number of evaluators drawn equals $floor(E_s)$

or the maximum number of drawn evaluators is reached. The process is still ensuring that no two evaluators from the same university are selected.

3. **Completing the draw.** Any remaining positions are filled from the entire pool of eligible evaluators, selected in random order until the total number of evaluators reaches the target max_draw. The process is still ensuring that no two evaluators from the same university are selected.

As with the 2012 draw, we replicate this algorithm to recover the ex ante draw probability of each potential evaluator, and each pair of evaluators.

B Alternative peer effects specification

In this section, we introduce an alternative specification that explicitly separates own-quality effects from peer effects. This approach has a key advantage. By isolating own quality from peer quality directly, it allows us to estimate own effects without relying on coefficient decompositions. This yields an easier interpretation of the estimates, as variation in peers' quality is orthogonal to changes in an evaluator's own quality by construction.

We estimate the following specification:

$$y_{i,j,e} = \beta_0 + \beta_1 T_j + \beta_2 T_{e,-j} + \beta_3 \mathbb{P}\left[drawn_j\right] + \beta_4 \mathbb{E}\left[T_{e,-j}\right] + \mathbf{X}_{i,j}\gamma + \varepsilon_{i,j,e}, \tag{B1}$$

where the outcome variable $y_{i,j,e}$ captures textual features of evaluator j's report for candidate i in exam e. 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. The term $\mathbb{E}\left[T_{e,-j}\right]$ captures the expected peer quality faced by evaluator j, conditional on being drawn:

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

where $\mathbb{P}\left[drawn_{k,e} \mid drawn_{j,e} = 1\right]$ 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 report writing, holding the quality of their peers fixed. The coefficient β_2 reflects the impact of average peer quality. Since committees consist of four members, $\beta_2/3$ identifies peer effects: it measures how a one–standard-deviation increase in the quality of a single peer affects the length and complexity of other evaluators' reports. Identification relies on random variation in peer composition around expected peer quality.

Table B1 report the estimation results from this specification. Consistent with the estimates from equation (3), we find no evidence that better-published evaluators write reports of different length or complexity. In contrast, we document robust peer effects of similar magnitude: being assigned a peer who is one standard deviation better published increases the length of evaluators' reports by about 7% of a standard deviation in column 2, corresponding to roughly 20 additional words per report.

Table B1: Committee composition and individual report writing (alternative specification)

	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 evaluators over the ten years preceding the exam. *Committee quality* (T_e) is the average quality-adjusted publication record of evaluators over the ten years preceding the exam. Similarly, *peer quality* ($T_{e,-j}$) is the quality-adjusted publication record of all other evaluators on committee, excluding evaluator j over the ten years preceding the exam. 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

Beyond increasing report length, higher-quality peers also prompt evaluators to write more complex reports. Specifically, being assigned a peer who is one standard deviation better published leads evaluators to use words with an *Average IDF* score that is approximately 6% of a standard deviation higher (column 3). Consistent with increases in both the length of reports and average complexity of words, the total complexity of reports also increases with peer quality (column 4).

In Table B2 we estimate a specification where we replace the continuous measure of peer quality with indicator variables for the number of *top researcher* peers on the committee. The point estimates suggest that there may be some non-linearity in the response to the number of top peers on the committee with respect to exerted effort, however our estimates are imprecise due to the lack of power.

Table B2: 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 2012 wave 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

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 CVs, 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 list a larger number of works on their CVs 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.

Table C1: Candidate research profiles in CV 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 2012 wave 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.

Next, we report the correlation between the measures coming from CVs 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.

Table C2: Candidate research profiles in CV 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 2012 wave 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.

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 4.1, 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.

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 2012 wave 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 2012 wave of evaluations. In columns 9 and 10 the sample is all candidates who applied for qualification in STEM+M fields in the 2012 wave of evaluations. Standard errors are clustered at the exam level.
*** $p < .01$, ** $p < .05$, * $p < .1$

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.

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

	STEM (1)	Overall (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 2012 wave of evaluations. In column 2, the sample is all candidates who applied for qualification in the 2012 wave of evaluations. Standard errors are clustered at the exam level.

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

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 CVs. 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

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

	Main-Text (1)	OpenAlex IRIS (2)	OpenAlex IRIS - Scaled (3)	Count High-Impact (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

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

	Main-Text (1)	Top Researcher (2)	AIS (3)	Articles WOS (4)	Articles (5)	Citations (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 Specification Tests

Table D1: 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
R^2	0.003	0.000	0.003	0.003	0.002

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 D2: First stage results from committee-level regression

	Final Committee Quality (1)
Initial Committee Quality	0.841***
	(0.075)
Candidate's Impact of Publication	0.000
	(0.000)
Candidate's Number of Publications	0.000
	(0.000)
Expected Committee Quality	0.298
	(0.277)
Observations	69020
KP rk Wald F-stat	125.395
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 D3: 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