Judging the Paper by Its Cover: Affiliation Bias in Conference Admissions*

Giacomo Gallegati[†] ⓡ Luca Favero[‡] ⓡ Águeda Solís Alonso[§] ⓡ Enrique Carreras[¶]

February 2025

Abstract

Connections predict career success across many occupations. In academia, conference participation is essential to build and maintain a professional network, thus contributing to academic success. In this paper, we experimentally test whether and how academic affiliation affects conference acceptance. We run a matched-pair field experiment leveraging the reviewing phase of an early-career workshop in economics. By randomly allocating anonymous papers with and without the submitting author's affiliation to matched pairs of reviewers, we find that affiliation disclosure leads to a substantial bias in favor of authors coming from prestigious institutions. Affiliation bias ultimately reduces the diversity of conference participants, particularly by lowering the representation of women and first-generation attendees. Importantly, we find that this bias is largely explained by in-group favoritism, as it is mainly driven by reviewers from similarly ranked institutions. Our results suggest that affiliation bias reinforces existing inequalities and hampers academic diversity.

Keywords: affiliation, bias, peer-review, conferences, early-career
JEL codes: A1, D8, I2, J71, C93

*The author order has been randomized following Ray ⓡ Robson (2018). We are profoundly indebted to the many early-career researchers who participated in the pilot and the final experiment. We are grateful for useful feedback and discussions to Mariana Blanco, Thomas Buser, Alessandra Casella, Pierluigi Conzo, Marina Della Giusta, Fulya Ersoy, Patricia Funk, Marc Gurgand, Jeanne Hagenbach, Nicolas Jacquemet, Andrea FM Martinangeli, Francesco Passarelli, Chiara Pronzato, Rafael Schütz, Giuseppe Sorrenti, Juan Vargas and participants of the 12th meeting of the Behavioral and Experimental Economics Network (BEEN), the Bocconi-Carlo Alberto-Cornell Workshop in Political Economy, the 2024 International Workshop of Applied Economics of Education in Catanzaro, seminars at PSE, UvA, University of Essex, University of Turin, University of Oviedo, Marche Polytechnic University and University of Insubria. Luca Favero acknowledges financial support provided by the ESRC Research Centre on Micro-Social Change (ES/S012486/1). We received ethics approval for this experiment from the Ethics Committee of Collegio Carlo Alberto and pre-registered it as AEARCTR-0012402.

[†]University Paris 1 Panthéon-Sorbonne, University of Turin and Collegio Carlo Alberto (email: giacomo.gallegati@carloalberto.org)

[‡]University of Essex and Collegio Carlo Alberto (email: l.favero@essex.ac.uk)

[§]University of Turin and Collegio Carlo Alberto (email: agueda.solisalonso@carloalberto.org)

[¶]University of Turin and Collegio Carlo Alberto (email: enrique.carreras@carloalberto.org)

1 Introduction

Connections and professional networks are key elements for career success across different occupations (Calvo-Armengol and Jackson, 2004; Cullen and Perez-Truglia, 2023; Lleras-Muney et al., 2020; Chetty et al., 2022). Academia is no exception: forming and maintaining relationships with peers and other researchers is an important element for developing a career in research. In particular, PhD supervisor's network predicts initial placement (Rose and Shekhar, 2023). In turn, personal connections through the PhD-granting institution and places of employment influence future citations (Head et al., 2019) and the probability of being promoted (Zinovyeva and Bagues, 2015), while co-author networks influence future productivity (Ductor et al., 2014) and funding success (Tsugawa et al., 2022).

Conferences serve as platforms for interaction and exchange of ideas, being a key driver for the visibility and success of future publications, particularly for less established scholars (Bellemare, 2022; Leite Lopez de Leon and McQuillin, 2020; Gorodnichenko et al., 2021). Conference presentations may even serve as measures of external recognition for tenure and promotion decisions (Chari and Goldsmith-Pinkham, 2017). As such, they play a pivotal role, not only in individuals' advancement, but also in the functioning of academia and the production of scientific knowledge.

Admission to academic conferences, similarly to publication, is primarily based on peer review. If these processes are not free from biases, there may be severe and unfair consequences for the careers of scholars. Unfortunately, multiple papers show how various personal characteristics and labels may influence the evaluation of research. In particular, previous studies have found that gender (Hóspido and Sanz, 2020; Samahita and Devereux, 2024; Card et al., 2020, 2022), seniority (Seeber and Bacchelli, 2017; Uchida, 2021), nationality (Tavoletti et al., 2022) or ethnicity (Pleskac et al., 2024), physical appearance (Hale et al., 2023), and author's prominence (Huber et al., 2022; Tomkings et al., 2017) might bias one's judgment.

In this paper, we conduct a field experiment to explore the role of affiliation bias in the peer review process of an early-career workshop in Economics at a French university. We gathered data from 140 early-career researchers who applied to the workshop and agreed to act as peer reviewers. We designed a matched-pair experimental design by creating couples of reviewers allocated to different treatment arms. Each couple consisted of a reviewer assigned to visible affiliation (VA) grading and another assigned to Non-Visible Affiliation (NVA) grading. We removed any identifying information in all papers assigned for peer review, including title, author identity, and acknowledgments. Papers assigned to VA reviewers retained author affiliation, while this information was removed for papers assigned to NVA reviewers. Importantly, our reviewers had no knowledge of taking part in an experiment during the grading phase and had no information about the experimental design.

Each couple received a block of eight randomly allocated papers from the pool of applications to the workshop. In our design, each reviewer only received papers prepared for either visible or Non-Visible Affiliation grading, but each paper was evaluated multiple times by reviewers across the two treatment arms. This design allows our preferred econometric specification to identify the causal effect of visible affiliation by exploiting variation within papers by block, keeping the quality of papers fixed, and taking into account that assessments might potentially be affected by the other papers in the block (the "grading on a curve" behavior described by Calsamiglia and Loviglio (2019)).

We collected data on grades, perceptions, and reviewer characteristics through a dedicated online survey. At the beginning of the reviewing phase, all applicants received a folder containing the papers they were assigned to review, instructions outlining the evaluation criteria, and a link to a survey for submitting their reviews. The grading criteria consisted of three aspects: relevance of the research question, quality of the research design, and quality of the writing. Reviewers were asked to score each paper from 1 to 10 on each of the three aspects above, to provide a suggestion for conference inclusion, and to express interest in discussing research with the author. We additionally collected data on various aspects of the reviewing process, including reviewers' views on the importance of signals when evaluating research and some demographic characteristics. Additionally, we elicited reviewers' perceptions of institutional quality by asking them to categorize institutions into different tiers based on perceived quality.

We find evidence that visible affiliation grading changes reviewers' evaluations and ultimately affects the composition of academic conferences in favor of top institutions. Reviewers assigned to VA grading gave a differential treatment to research from prestigious institutions across all our grading criteria. In particular, we document a differential treatment effect for papers from top institutions. These papers gain 1.62 points in the Overall score, out of 30, when displaying their affiliation, in addition to the small and statistically insignificant, effect for papers from lower-ranked institutions, which lose 0.48 points when they show their affiliation. These estimates correspond to a total premium of 1.14 points for top institutions. This effect is reflected across all grading criteria. The differential treatment effect maps into a differential improvement of 18.59 positions for highly ranked institutions when displaying their affiliation, corresponding to 11.9% of the total number of applications in the ranking. We next document through a simulation exercise how affiliation bias shapes the composition of conferences in favor of prestigious institutions across a wide range of acceptance rates. Applying to our sample the acceptance rates from the annual meetings of the American Economic Association and the European Economic Association,[1] we find that affiliation bias increases the acceptance likelihood of papers from top institutions by 12.9 and 18.2 percentage points, respectively, compared to papers from lower-ranked institutions. Importantly, we also find affiliation bias to play an effect on conference composition in terms of the gender and socio-economic background mix of the attendants. We show that under the NVA scenario, the representation of females and first-generation authors would be virtually always higher compared to VA grading. This

[1]The 2024 ASSA Annual Meeting had a 13% acceptance rate, while the 2003 European Economic Association meeting had a 43% acceptance rate.

result is likely driven by the different sociodemographic composition across different types of institutions and possibly by the different application behavior of females in non-elite institutions (Coffman et al., 2024; Farré and Ortega, 2024). In addition, we provide evidence suggesting that affiliation bias is driven by papers of relatively lower quality. Using NVA grades and predicted citations as proxies for submission content, we show that affiliation bias might also reduce conferences' overall quality. We then assess the external validity of our findings by asking a large language model (LLM) to simulate our results. We do so, by providing the LLM with a brief description of the design, sample, and context of the experiment, and eliciting predictions on the results for Overall score and Suggestion for inclusion. We show that the model can closely approximate our findings. Next, we replicate the exercise by requesting estimates for a sample of senior economists from prestigious institutions. The LLM effectively distinguishes between the two samples, producing estimates consistent with our original findings.

Furthermore, we explore the mechanisms underlying our results and find evidence linking affiliation bias to reviewer characteristics and different types of discrimination. We first show that affiliation bias is driven by reviewers from more prestigious universities, suggesting the presence of in-group favoritism (Akerlof and Kranton, 2000; Chen and Li, 2009) or *club-like* behavior as described by Carrell et al. (2024) in publishing. Finally, we show how part of affiliation bias is based on beliefs on differences in quality, statistical discrimination (Arrow, 1973), and on a differential preference towards papers from prestigious institutions, as in taste-based discrimination (Becker, 1957).

Our study contributes to the ongoing debate on fairness and cognitive biases across review stages in academia. Over the past years, a series of research works have shown how biases and unequal treatment of some minorities and collectives (Card et al., 2020, 2022; Huber et al., 2022; Pleskac et al., 2024) may contribute to the lack of diversity in academia. However, affiliation bias has received less attention from a research standpoint, and consequently, the discussion about its implications and possible solutions is currently at a much less developed stage.

Nevertheless, existing evidence (Ersoy and Pate, 2023), highlights how a part of the gender bias in the publication process can be related to the different gender composition of top and non-top Economic departments. Moreover, Schultz and Stansbury (2022) find for the US context a marked correlation between university status and lack of socioeconomic diversity. We document a similar pattern in our study looking at parental education, as shown in Table A.1 in Appendix A.1. These observations underline how affiliation bias can shape and amplify other inequalities and ultimately reduce intellectual diversity in the discipline (Bayer and Rouse, 2016). We focus on Economics, a discipline marred by strong inequalities, concerns of elitism, underrepresented minorities, concentration of talent, and with a large "publication gap" between top and lower-ranked institutions (Fourcade et al., 2015; Freeman et al., 2024; Hoover and Svorenčík, 2023; Schultz and Stansbury, 2022). Taken together, these features suggest that affiliation may hold higher relevance in Economics compared to other fields. Previous studies on affiliation bias in academia have mostly focused on publication in top journals and citations (Blank, 1991; Ersoy and Pate, 2023). Evidence of affiliation bias in conferences is scarce, with the notable exception of Uchida (2021), and non-existent for conferences in Economics or a population of early-career researchers. To the best of our knowledge, we are the first to show that differential treatment based on author's affiliation is not exclusive of senior, established scholars, but a pervasive behavior within the discipline that starts at the doctoral stage. Moreover, with the exception of Ersoy and Pate (2023), previous studies have mostly focused on the impact of double-blind reviewing, and cannot tease out the effect of different elements of the author's identity from the overall impact of double-blind review (Blank, 1991; Pleskac et al., 2024).

This experiment, by randomly showing or concealing affiliation to reviewers while keeping the rest of the information hidden across treatment arms, is specifically designed to isolate the causal effect of affiliation on the evaluation of research works. The context of our experiment is particularly well suited to alleviate concerns of non-compliance with anonymization. Previous works leveraged a sample of established reviewers and research that was either available online or potentially recognizable by a non-negligible proportion of reviewers (Ersoy and Pate, 2023; Blank, 1991). We focus on work-in-progress research of junior scholars, and a population of junior reviewers, making it less likely that the papers are available online, or recognizable by peers. We show causal evidence that, *ceteris paribus*, researchers from top institutions benefit from signaling their affiliation, and that affiliation bias affects the composition of academic conferences in favor of more selective institutions. Our findings document the existence of widespread biases that arise in the first stages of academic careers, through access to conferences, hampering publication potential and perpetuating inequalities in academia.

The remainder of this paper is organized as follows. Section 2 describes the context of our experiment, and Section 3 presents our empirical strategy. Section 4 illustrates the main results, Section 5 reflects on potential mechanisms, Section 6 discusses robustness checks, and finally Section 7 concludes.

2 Context and experimental design

We designed and implemented a field experiment during autumn 2023, leveraging the application and review phase of the PhD Workshop in Networks and Political Economy at the University Paris 1 Panthéon-Sorbonne. This conference accepted extended abstracts from PhD candidates and postdocs, focusing on political economy but welcoming applications from all fields and methodologies in economics.

To be eligible for the workshop, applicants had to agree in advance to review 8 papers from fellow early-career researchers. Reviewers were not informed that they were part of an experiment during the reviewing process. This requirement facilitated the organization of the workshop, by enabling peer review and hence reducing the reviewing burden on the organizers, while also playing an important role in our experimental design. Firstly, it allowed us to expand the number of reviewers to a larger sample than we could have achieved by relying solely on the conference organizers. Secondly, it provided a strong incentive for each submitter to complete the review process, enabling us to collect data with a high response rate without alerting reviewers about the experiment. Third it allows us to observe the identity of both submitters and reviewers, a feature we use to explore heterogeneity.

We designed a matched-pair experimental design by creating couples of reviewers allocated to different treatment arms. We stratified reviewers according to the selectivity of their institution. We classified each institution as selective or non-selective based on their presence in the top 100 of the 2023 QS subject-specific Ranking for Economics and Econometrics.[2] We received 156 papers submissions, 58 came from selective institutions and 98 from non-selective institutions. For each stratum we randomly assigned reviewers to either visible or Non-Visible Affiliation grading with equal probabilities. We made sure to allocate applicants from the same institution to the same treatment arm, minimizing concerns of potential treatment spillovers. This feature of the design was made possible by randomizing treatment at the level of each affiliation. We then created pairs by randomly matching selective with non-selective reviewers that were allocated to alternative treatment arms. The remaining non-selective reviewers were matched together in couples where both reviewers came from non-selective institutions. Figure 1 below visualizes our experimental design. The stratification ensures that treatment is balanced across the institutional quality of the reviewer and regulates how couples are formed. This design allows us to explore heterogeneous treatment effects according to the selectivity of reviewers by comparing reviewers from the same stratum allocated to different treatment arms.[3]

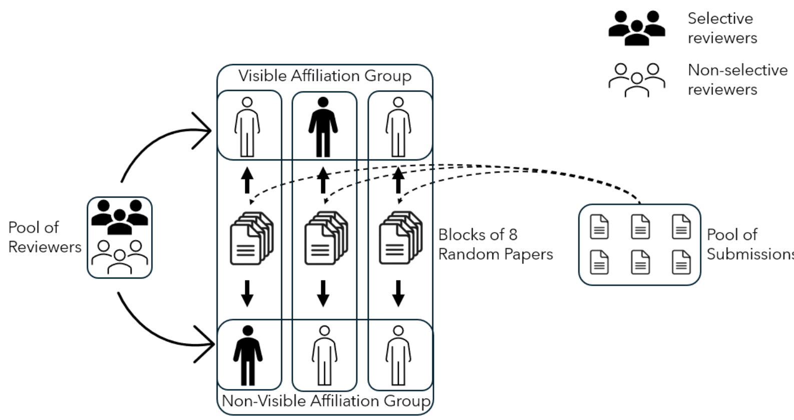
Each pair received a block of eight randomly allocated papers, which differed only in affiliation visibility. Before sending it to the reviewers, we modified each paper into a visible (VA) and Non-Visible Affiliation (NVA) version. In both versions, title, acknowledgments, author's name, indications of preliminary work, and any potential identifying information were removed. However, the VA version retained the applicant's affiliations, which was instead removed from the NVA one.[4] Next, we converted all files into non-searchable PDFs to increase the cost of searching for the missing identifying information online. Finally, we assigned each file a randomly generated name for each treatment arm to prevent reviewers from identifying any characteristics of the paper, such as the title or author, based on the file name. This approach mitigated concerns about potential communication between

[2]We also added to the list of selective institutions two additional establishments that are not present in the ranking but had clear standing in economics. The full list of establishments classified as selective institutions according to the stratification variable can be found in Appendix A.4

[3]FigureA.2 in Appendix A.1 reports the balance check of reviewer characteristics

[4]For authors with multiple affiliations, we kept all affiliations indicated during paper submission. Any co-authors' affiliations were suppressed. Appendix A.2 reports an illustrative example of a paper across its two versions

Figure 1: Experimental design



reviewers and the risk of them becoming aware of the different treatment arms.

Each reviewer in each pair was either assigned to VA or NVA grading, and had no knowledge of the presence of a matching reviewer assigned to the alternative treatment arm. Importantly, while reviewers only saw papers with visible or hidden affiliation, each paper was assigned to multiple reviewers in each treatment arm. This procedure allows us to explore variation within paper by blocks, keeping the relative quality of each paper fixed and accounting for grading on a curve (Calsamiglia and Loviglio, 2019).

Reviewers received an email with a link to a data collection tool hosted on LimeSurvey, where their grades and additional information were collected. To ensure the highest level of compliance, various reminders were sent to reviewers before the deadline. This approach, combined with the requirement to serve as reviewers to be considered eligible for the workshop, resulted in a high response rate, with a sample of 140 reviewers, producing a total of 1,120 reviews.[5]

During the review phase, applicants received a Google Drive folder containing the documents they were assigned to review, along with instructions that outline the evaluation criteria and a link to a survey to submit their reviews.[6] The grading criteria consisted of three aspects: relevance of the research question, quality of the research design, and writing quality. Reviewers were explicitly informed that all grades would be used for determining acceptance into the conference, with no distinction on their importance. Reviewers were asked to score each paper from 1 to 10 on each of the three aspects above and to provide a suggestion for conference inclusion. Additionally, reviewers were asked to express interest

[5]We received a total of 156 applications and 140 replies amounting to a response rate of 89.7% balanced across the two treatment arms.

[6]Refer to Appendix A.2 for more details

in discussing research with the author.

After submitting their grades, reviewers encountered a new section of the survey, designed to test for potential experimenter demand effect and treatment salience that may have affected reviewers' behavior. Experimenter demand effect is unlikely to be a major concern in our setting, given that participants were unaware of the experiment. Regardless, we leveraged the information collected in this section to run a robustness check and rule out any remaining concerns.

Subsequently, reviewers were asked to assess the importance of signals when evaluating a paper, and, further into the survey, they were tasked with categorizing institutions into different tiers based on perceived quality. The tiers were: "elite", "above average", "average", "below average", and "I don't know". The institutions they had to rank included those affiliated with the authors of the papers under review, the reviewers' institution, and the one of the reviewer they were matched with. Finally, in the last section, we collected socio-demographic information about the participants.[7]

We took particular care in our experimental design to minimize the risk of reviewers finding out about the experiment. First, reviewers were told they were part of an experiment after submitting the grades when providing socio-demographic characteristics. We also made sure that no reviewer would be provided with a paper from their university or with a co-author to avoid any casual talking about the reviewing process, which might have put at risk the compliance with treatment allocation, and thus the validity of the experiment. Importantly, the online survey did not allow to navigate back to previous, already completed, sections. This last precaution is particularly relevant as reviewers were informed about the collected data being used for research purposes only in the last section of the survey, after all the outcomes were collected and could not be amended. To this end, we designed the survey across separate sections, not allowing reviewers to alter their responses as they progressed throughout the different stages of the data collection, mitigating possible concerns of manipulation of previous answers and grades.

3 Empirical strategy

This paper exploits the random allocation of reviewers to visible and Non-Visible grading and variation in affiliation across papers to identify the presence of affiliation bias in the peer-review process determining acceptance to an early-career conference in economics.

Our main specification is described in Equation 1 below:

$$Y_{ijb} = \alpha + \beta_1 \text{VA}_j + \beta_2 \text{VA}_j \times \text{Top75}_i + \boldsymbol{\beta_3'}\mathbf{X}_j + \gamma_{ib} + u_{ijb} \qquad (1)$$

Y_{ijb} measures the outcome of interest for paper i evaluated by reviewers j in block b. VA_j is a dummy equal to one if a paper is evaluated by a reviewer in the visible affiliation

[7]More information regarding the questions collected during the survey can be found in Appendix B.1

treatment arm. $\text{VA}_j \times \text{Top75}_i$ is defined as a dummy equal to one for papers coming from a Top 75 institution which are evaluated by reviewers in the visible treatment arm.[8] β_2 is the coefficient of interest measuring any differential response to visible affiliation grading for top institutions. We refer to this differential effect as evidence of "affiliation bias". $\mathbf{X}_j$ is a vector of reviewer characteristics measuring gender, year in the PhD, country of residence, parental education as a proxy for socio-economic background, and our stratification variable for selective reviewers. γ_{ib} are paper-by-block FEs. We cluster standard errors at the reviewers' institution level, mirroring treatment assignment (Abadie et al., 2023).[9]

We define "Top" institutions in our variable of interest as those in the top 75 positions of the QS subject-specific ranking for Economics and Econometrics. The choice of a top 75 cutoff comes from evidence collected in the survey where we ask reviewers about their perceptions of elite institutions. We find a strong discontinuity in the proportion of reviewers considering an institution as "elite" beyond the Top 75 cutoff once we plot the subjective perceptions of university quality against the QS ranking, the objective measure of institutional quality we leverage in this paper.[10] Due to the context of our experiment, taking place in a new early-career conference in Europe, our sample is less selective than in related papers (Ersoy and Pate, 2023; Uchida, 2021), which makes us choose a lower threshold to characterize top institutions. Hence, we believe that our results might be a lower-bound estimate for the effects that could be found for more prestigious institutions.[11]

Unbiased estimation of β_2 requires treatment status to be orthogonal to reviewer characteristics and the quality of each submitted work. Random assignment of reviewers to visible and Non-Visible grading guarantees that reviewers are balanced on both observable and unobservable characteristics. We provide evidence that reviewers are similar across all individual characteristics we collected in Appendix A.1.1. We also made sure that all reviewers applying from the same institution shared the same treatment arm to minimize concerns over treatment spillovers. The inclusion of paper-by-block FE ensures that identification of β_2 comes from comparing the same paper, from the same block, read by a reviewer in the VA treatment arm versus a reviewer assigned to NVA treatment. This comparison keeps quality fixed while being robust to grading on a curve.

We modify our main specification when analyzing treatment effects by the selectivity

[8]Throughout the paper we use the terms "selective" and "non-selectives" to refer to the stratification variable and "Top75" when talking about the institution of the papers assigned to reviewers.

[9]There are 90 clusters in our dataset. The inclusion of paper-by-block FEs in our main specification captures variation within the couple and ensures that standard errors do not need further adjusting to account for the pairwise nature of our experimental design as discussed in de Chaisemartin and Ramírez-Cuéllar (2024).

[10]See Figure **??** in Appendix A.3

[11]In Figure A.2 in Appendix A.6, we explore the robustness of our results to different cutoff choices. In Table A.7 in Appendix A.6 we also consider an analysis using individual reviewers' assessments of the institutions to define our top category. Results remain quantitatively similar, but we decide not to use these as our main estimates because of concerns about this classification being affected by treatment status.

of reviewers in Section 4.3 to take into account the specificities of our experimental design. As discussed in Section 2, our experiment leverages couples of matched reviewers where for each couple we have either a selective reviewer matched with a non-selective reviewer or both non-selective reviewers matched together. For this reason, we explore heterogeneity by reviewer type by changing our fixed effects from paper-by-block to paper. This new specification still keeps the quality of each paper constant and identifies our coefficient of interest by comparing reviewers who were assigned to alternative treatment arms and who read papers from different institutions.

4 Results

4.1 Main results

Table 1 reports the estimates for equation 1 using as dependent variables the paper's Overall score (the sum of the three grades described in Section 2), the change of positions in the ranking of the paper by Overall score, and a dummy indicating whether the paper was suggested for inclusion by the reviewer.

Table 1: Main estimates

	(1) Overall score	(2) Position in ranking	(3) Suggestion for inclusion
VA	-0.48	2.87	0.08*
	(0.55)	(2.11)	(0.04)
VA × *Top 75*	1.62***	18.59***	0.20***
	(0.55)	(3.42)	(0.06)
N	977	985	985
Control mean	20.31	78.87	0.55

Note: Dependent variable for "Overall score" refers to the sum of grades for research question, research design and writing scores. "Position in ranking" is computed by taking the rank of "Overall score" for each paper separately by treatment arm. "Suggestion for inclusion" is a dummy equal to 1 combining evaluations for "Definitely accept" and "Probably accept" and zero for "Maybe accept" and " I think this paper should not be accepted". The control mean refers to the average outcome in NVA grading. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin, a stratification dummy for selective reviewers and paper-by-block FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

Disclosing affiliation has mostly small and non-statistically significant effects for papers from universities not in the Top 75. In contrast, our estimates show a differential positive effect of 1.62 points (out of 30) on the total score awarded by VA reviewers to papers from

Table 2: Disaggregation of the effect on Overall Scores

	(1) Overall score	(2) RQ score	(3) Design score	(4) Writing score
VA	-0.48	-0.14	-0.27	-0.04
	(0.55)	(0.18)	(0.21)	(0.20)
VA × *Top 75*	1.62***	0.35*	0.65***	0.60***
	(0.55)	(0.19)	(0.24)	(0.19)
N	977	981	979	979
Control mean	20.31	6.96	6.62	6.72

Note: Dependent variable for "Overall score" refers to the sum of grades for research question, research design and writing scores. Scores range from 1 (lowest) to 10 (highest). The control mean refers to the average outcome in NVA grading. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin, a stratification dummy for selective reviewers and paper-by-block FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

Top 75 institutions.[12] In Table 2 we delve deeper into this premium by breaking down the overall increase into the three grading criteria, and we find that the effect is equally driven by the writing clarity and the research design scores, with a smaller positive effect on the research question score as well.

This premium is reflected in a marked increase in the probability of proposing the paper for inclusion into the conference, and has strong implications for the actual chances of admission. With respect to the other papers, submissions from Top 75 institutions gain on average 18.59 positions in the ranking of total scores when their affiliation is revealed, corresponding to almost 12% of the total number of applications.

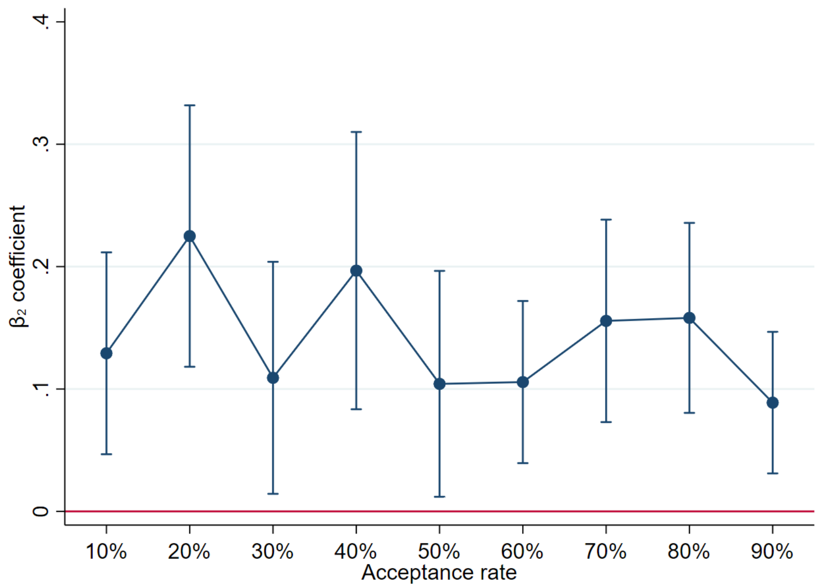
Ultimately, effects on probability of acceptance do not only depend on the reviewers' engagement in bias, but also on the characteristics of the pool of candidate papers (composition in terms of top/non-top applicants, distribution of grades within and between both groups, etc.) and the conference acceptance rate. For this reason, in Figure 2 we perform a simulation analysis on the probability of being accepted across different thresholds. A possible concern for our simulation exercise is that application and reviewing behavior might not be the same across different expected acceptance rates. However, because our workshop is a newly established event, with no information on previous or expected acceptance rates, we believe this type of concern is unlikely to play a major role in our context.

We find qualitatively similar results for acceptance rates between 10% and 90%, with an increased probability for papers coming from Top 75 institutions ranging between 9 pp to 22 pp compared to non-top affiliated papers. To put these rates into perspective, if the acceptance rate had been 13% as in the 2024 ASSA Annual Meeting of the American

[12] We report p-values adjusted for multiple hypothesis testing in Table A.8 in the Appendix

Economic Association[13] or in the 2017 NBER Summer Institute,[14] we would have found a visible affiliation differential for Top 75 institutions of 12.9 percentage points in our sample, whereas if it had been 43%, as the 2003's EEA Meeting,[15] it would have been 18.2.

Figure 2: Affiliation bias by acceptance rate



Note: Figure reports 95% confidence intervals for separate estimates of β_2 from Equation 1. Dependent variable refers to a dummy taking value 1 for papers accepted according to different acceptance rates. Estimates come from separate models, one for each acceptance threshold. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin, a stratification dummy for selective reviewers and paper-by-block FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level.

The affiliation premium on the probability of being accepted at the paper level has profound implications on the final composition of the event. Figure 3 shows the proportion of Top 75 applications that would get hypothetically selected for the conference under the Non-Visible Affiliation grading scenario, and the difference with the visible affiliation scenario. Because of the expected correlation between institutional rank and research quality, Top 75 institutions would be over-represented in the conference across virtually all acceptance rates even if only Non-Visible Affiliation reviews were considered. Nevertheless, this over-representation gets magnified by affiliation bias, which causes the substitution of

[13]Source: `https://www.aeaweb.org/econharmony/`

[14]Source: Chari and Goldsmith-Pinkham (2017)

[15]Source: `https://warwick.ac.uk/fac/soc/economics/staff/academic/walker/eea-es_joint_meeting.pdf`

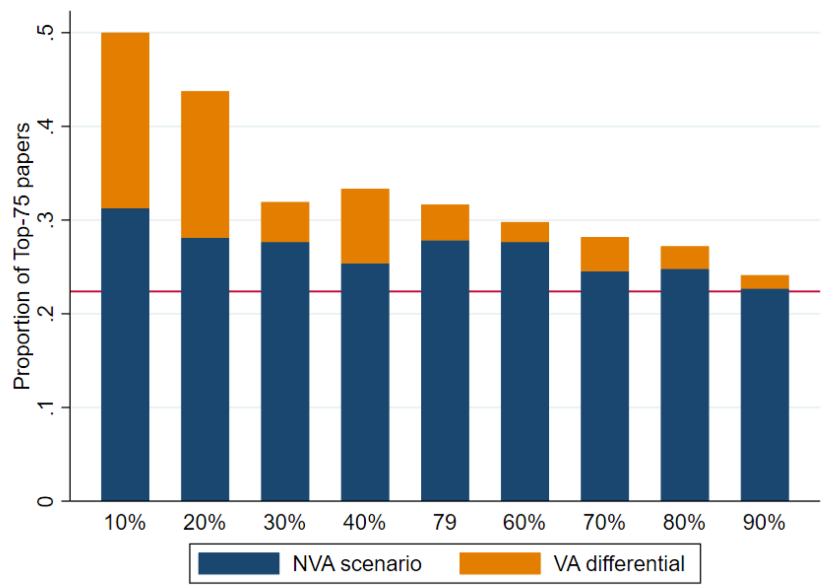
non-top papers with high NVA scores with Top 75 papers with similar, but less positive, reviews. The magnitude of this phenomenon is sizable, with the VA differential displayed in Figure 2 ranging from 1.42 pp in the unlikely scenario of a 90% acceptance rate, to 18.75 pp and 15.63 pp for the more realistic acceptance rates of 10% and 20%.

Moreover, we show that affiliation bias in conference composition also affects the gender and socio-economic background of attendees. Replicating the previous analysis on the proportion of females and people whose parents do not have a university degree, we find that under the Non-Visible Affiliation scenario, the representation of these two collectives would be virtually always higher. This result is likely driven by the different sociodemographic composition in both types of institutions (see Table A.3 and Schultz and Stansbury (2022)), and possibly by different application behavior of females in non-Top 75 institutions.[16] Our findings mirror Ersoy and Pate (2023)'s results on the role of department composition on the publication gender gap, proposing affiliation bias as an additional channel behind the barriers to the early career development of these two groups.

Our results, coming from a non-US, new conference targeted at junior researchers, are not directly comparable to those of the existing literature because of the differences in the experimental setting, the pool of papers, and the characteristics of reviewers. On the one hand, the sample of universities is not as selective as the universities of Blank (1991), Ersoy and Pate (2023) and Uchida (2021). For this reason we set a lower threshold to define top institutions, which could push the estimates towards zero. On the other hand, affiliation could be a more salient signal in our population, due to the lack of alternative sources of information (e.g. name of the authors), and the inexperience of reviewers, who may need to rely more on the available signal (Spence, 1973). If the latter were true, we may expect larger estimates of bias. Despite these differences, our estimates are qualitatively similar to those of Blank (1991), Ersoy and Pate (2023), and Uchida (2021), confirming that visible affiliation grading leads to differential treatment between top and non-top institutions, also in this context.

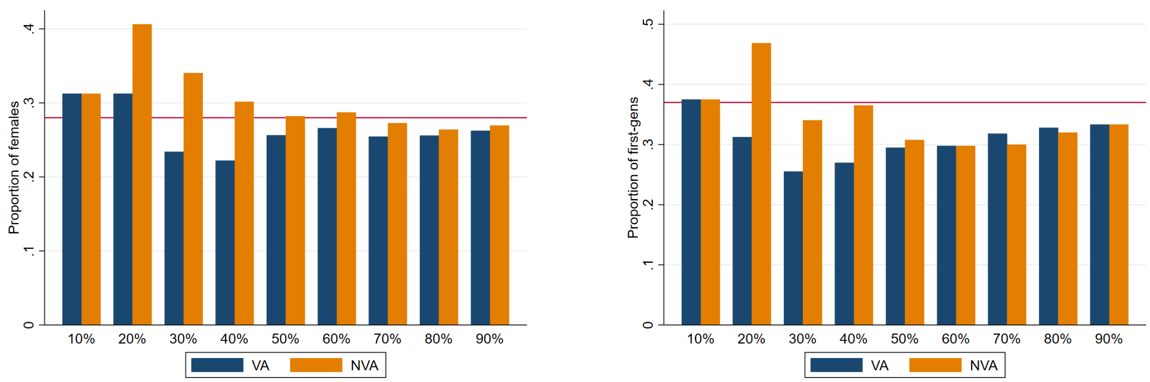
[16] A body of research has found that females are significantly less likely than equally qualified males to apply to challenging work (Coffman et al., 2024) or academic (Farré and Ortega, 2024) opportunities, especially in settings perceived as competitive (Flory et al., 2015) or with ambiguous requirements (Coffman et al., 2024). Following this research, our hypothesis is that females affiliated with less prestigious institutions might tend to feel more under-qualified with respect to other applicants, and apply to conferences only with relatively higher-quality submissions.

Figure 3: Conference composition by acceptance rate



Note: The blue bars report the proportion of hypothetically selected papers for each acceptance rate under the Non-Visible Affiliation grading scenario that were produced by a Top 75 institution. This proportion is computed by ranking papers by their average NVA Overall score received, taking the correspondent top %, and computing how many of them were submitted by a participant from this group. The orange bar shows the difference between the results obtained by this exercise and the analogous one using only reviews from the VA treatment group. The red horizontal line displays the proportion of Top 75 papers in the full sample, 22.38%.

Figure 4: Socio-demographic conference composition by treatment arm and acceptance rate



(a) Gender

(b) Socio-economic background

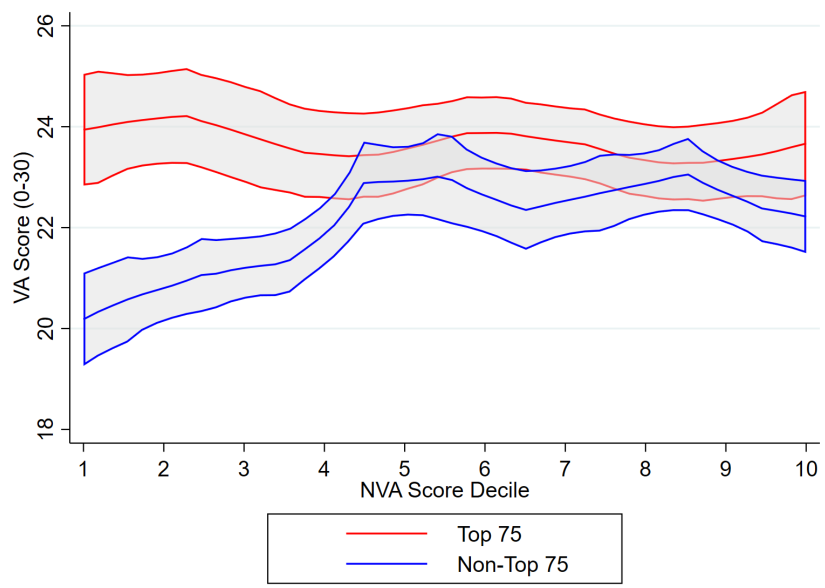
Note: Bars report the proportion of hypothetically selected papers for each acceptance rate under each grading scenario that were submitted by a member of a minority. This proportion is computed by ranking papers by their average Overall score received under both treatment arms, taking the correspondent top %, and computing how many of them were submitted by a submitter of the correspondent characteristic. The red horizontal lines display the proportion of papers submitted by these groups in the full sample, which amount to 28% for females, and 37% for first-gens.

4.2 Effects of blinding on selection quality

While our estimates provide suggestive evidence that NVA grading produces a more diverse conference composition in terms of individual characteristics and institutional rank, there might be concerns that this comes at the expense of the quality of the selected works. If there is a statistical association between the reputation of the author's institutional affiliation and submission quality, concealing this information could harm the reviewers' capacity to evaluate research.

Figure 5 displays the relationship between Non-Visible Affiliation and Visible Affiliation Scores. For papers not coming from a Top 75 institution there is a positive relationship between the paper's NVA Score decile and its VA Score, showing that better papers get higher VA Scores. For papers coming from Top 75 institutions we observe a different pattern. For this group of papers, the relationship between NVA and VA Scores seems to be almost flat, revealing that Top 75 papers receive similar VA grades regardless of their content, and suggesting that the *affiliation premium* we document in the previous subsection might be driven by papers at the lower end of the quality distribution.

Figure 5: Relationship between VA Scores and paper quality



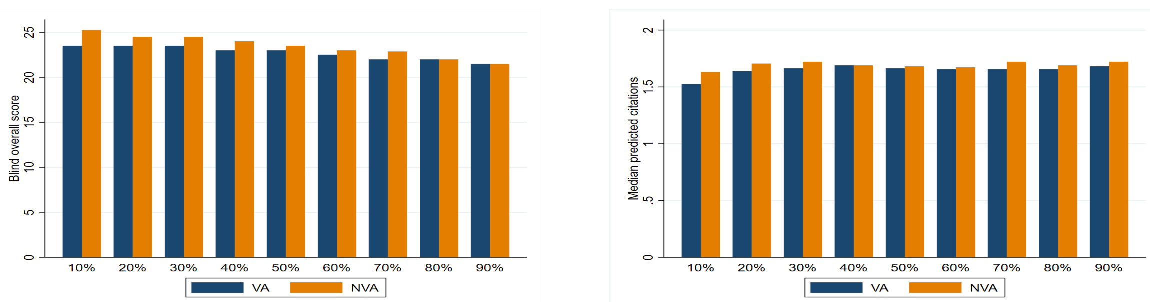
Note: Figure shows the relationship between Visible Affiliation scores and paper quality by paper's institution status, measured by the decile rank in Non-Visible Affiliation scores, using a kernel-weighted local polynomial regression. Shaded areas correspond to 95% confidence intervals.

These findings suggest that NVA grading should not decrease the average quality of selected submissions as it primarily impacts lower-quality works, which would be replaced by higher-quality papers from Non-Top 75 institutions.

The objective quality of a submission is difficult to measure. Related research has usually employed data on publication or citations to this end (Uchida, 2021). Although this information is not available for our sample, given that it consists of mostly preliminary works, we circumvent this data limitation by proposing two proxies. First, the scores received under Non-Visible Affiliation grading, which are exempt of the possible influence of biases as they are given by reviewers who graded completely anonymized papers. Second, we follow Iaria et al. (2024)'s approach and make use of Schwarz (2023)'s procedure to obtain predictions of log citations for each submission based on its title. The procedure is based on a text regression model, which is trained on a dataset of paper titles and citations from Clarivate Web of Science covering the years 1900 to 2010. As such, this measure is more a reflection of topic popularity than overall paper quality, but offers the advantage of not being directly affected by the existing biases in the publication process discussed by Card et al. (2020) or by biased citation patterns (Lawson, 2023).[17]

We present in Figure 6 the results of our simulation exercise on the mean quality of the selected works for each acceptance rate by treatment arm, using both of our proposed measures of quality. Both figures reveal the same pattern: under all acceptance rates studied, the average quality of the accepted submissions is equal, if not higher, under NVA grading.

Figure 6: Quality of selected works by treatment arm and acceptance rate



(a) NVA Scores

(b) Predicted citations

Note: Bars report the median quality of papers selected for the conference for each acceptance rate under each grading scenario. These figures are computed by ranking papers by their average Overall score received under both treatment arms, taking the correspondent top %, and computing the median of the corresponding measure of quality. We report this exercise using two proxies for quality of the works: the score received under Non-Visible Affiliation grading (on the left figure) and the predicted citations (right) using the method developed by Schwarz (2023); Iaria et al. (2024).

From these two exercises, we conclude that, at least for our sample, the advantages of Non-Visible Affiliation grading for conference composition diversity do not come at the

[17]Nevertheless, this measure could be affected by some *indirect* types of bias if, for example, authors from lower-ranked institutions in our sample do research in less cited fields or choose titles with terms that have been historically associated with less citations.

expense of a lower quality of selected submissions. This result suggests that the information conveyed by author's affiliation does not improve the reviewers' judgment of paper quality, in line with Pleskac et al. (2024).

4.3 Reviewer heterogeneity

In a further analysis, we explore the possibility of heterogeneous responses to treatment, according to the quality of the institutional affiliation of the reviewer. To do so, our main specification (1) needs to be modified in two ways. First, we use our stratification to divide our sample into "selective reviewers" (those belonging to a Top 100 institution according to the 2023 QS Economics and Econometrics ranking) and "non-selective reviewers". Secondly, our main model contains paper-by-block fixed effects, which vary at the couple level, as discussed in Section 3. However, because some couples in our sample are composed by a selective and a non-selective reviewer, maintaining these fixed effects precludes the identification of our coefficients of interest when dividing the sample. We substitute paper-by-block fixed effects with paper fixed effects, which still keep the quality of the paper fixed, while allowing us to explore reviewers' heterogeneity. Reassuringly, our main estimates remain virtually unaffected by this change of FEs, as shown in Table A.7 in the Appendix, making the results of this section comparable with our main estimates.

We find evidence that the score premium for Top 75 universities' papers is mainly driven by reviewers from similarly ranked universities, consistent with the existence of a systemic affiliation bias or in-group favoritism among these establishments, in line with Akerlof and Kranton (2000); Carrell et al. (2024); Bethmann et al. (2023); Reingewertz and Lutmar (2018). In particular, affiliation bias on Overall score goes from 1.09 for the whole sample to 3.15 points for selective reviewers. The differential effect on the probability of being suggested for inclusion also increases from 16 to 27 percentage points. In terms of position in the ranking, this increase translates into gaining 24 positions, compared to the 10 of our full-sample results. Results for the non-selective sample are closer to zero and not statistically significant, with the exception of the differential effect for the probability of being suggested for the conference, which is nevertheless about half the magnitude of what we find for selective reviewers.

Table 3: Heterogeneity: By reviewer type

	(1) Overall score	(2) Position in ranking	(3) Suggestion for inclusion
Panel A: Selective Reviewers			
VA × Top 75	3.15**	24.03***	0.27**
	(1.17)	(8.56)	(0.12)
Paper FE	Y	Y	Y
Paper by Block FE	N	N	N
Observations	396	398	400
Panel B: Non-Selective Reviewers			
VA × Top 75	0.62	5.44	0.14*
	(0.74)	(5.70)	(0.08)
Paper FE	Y	Y	Y
Paper by Block FE	N	N	N
Observations	700	701	701

Note: Dependent variable for "Overall score" refers to the sum of grades for research question, research design and writing scores. "Position in ranking" is computed by taking the rank of "Overall score" for each paper separately by treatment arm and reviewer type. "Suggestion for inclusion" is a dummy equal to 1 combining evaluations for "Definitely accept" and "Probably accept" and zero for "Maybe accept" and "I think this paper should not be accepted". The control mean refers to the average outcome in NVA grading. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin and a stratification dummy for selective reviewers. Panel A refers to the subsample of selective reviewers while Panel B refers to non selective reviewers, both panels control for paper FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

4.4 External validity - a LLM approach

In this subsection, we assess the external validity of our estimates. Recent research has begun investigating whether large language models can provide insights into social science experiments by predicting their outcomes (Horton, 2023). This approach mimics the more traditional approach of predicting experiments by human experts (DellaVigna et al., 2020, 2019). Large language models have been proven effective in forecasting results from both lab (Aher et al., 2023; Lippert et al., 2024) and field experiments (Chen et al., 2024). Notably, Chen et al. (2024) found that the highest prediction accuracy in field experiments is achieved when the model is provided with a clear description of the design, context, and characteristics of the sample before eliciting its forecast. This approach closely aligns with Lippert et al. (2024), who successfully replicated results from various experiments with high precision. Their simulations matched experimental results with an accuracy of 0.89, slightly outperforming human experts (0.87). We follow Lippert et al. (2024) to develop two prompts we feed to a large language model. In the first prompt, which we refer to as "RCT prompt", we provided a detailed description of our experimental design, including information on the conference and the descriptive statistics of applicants who applied to our conference. We then asked the model its best numeric predictions for β_1 and β_2 coefficients for Overall score and Suggestion for inclusion, the key outcomes from Table 1,[18] along with their expected significance levels. In the second prompt, "Out of sample prompt", we dropped any reference to the conference being targeted at early-career researchers and removed any descriptive statistics of what the sample looks like. We specified instead that both reviewers and applicants were senior economists from leading U.S. and European economics departments, and are typically appointed to the editorial boards of prestigious economics journals. We designed this second model to compare predicted treatment effects from a sample of junior economists with results from a sample of selective and senior academics. As in the first prompt, we then ask for its best predictions for β_1 and β_2 coefficients for Overall score and Suggestion for inclusion.[19]

For this simulation, we used the then-current ChatGPT model, gpt-4o-2024-08-06,[20] with the default temperature setting of 1. The task was repeated 1,000 times, each with a different random seed to ensure variability.[21] To improve replicability, we ensured that the same set of seeds would be used if the simulation were rerun. We report the results of this exercise in graphical form below. In each graph, we plot the density of simulated coefficients from the "RCT prompt", which replicates our experimental conditions, and of the "Out of sample prompt" that extends our design to a setting of senior scholars. We superimposed the value of our estimated coefficients as a vertical line and shaded the portion of the density that falls within the 95% confidence interval of our coefficient.

Overall, the simulated coefficients from the "RCT prompt" fall close to our estimates. Figure 7 reports results for Overall score in Panel A and for Suggestion for inclusion in Panel B. For both β_1 and β_2, the mass of the simulated distribution lies close to the estimated coefficient and well within the range of our 95% confidence intervals.

Similarly, we assess what the simulated results would be for a conference where both applicants and reviewers are senior, top-tier academics.[22] The results of this second simulation are consistent with our discussion in Section 4. Specifically, as discussed above, reviewers from more prestigious institutions tend to exhibit greater bias, in line with Chat-

[18]We do not ask the model to predict coefficients for Position in ranking because those coefficients are highly dependent on the number of applications and individual scores.

[19]We report both prompts in Appendix A.5.

[20]The knowledge cutoff of the model is October 2023, prior to our experiment. Source: `https://platform.openai.com/docs/models#gpt-4o`

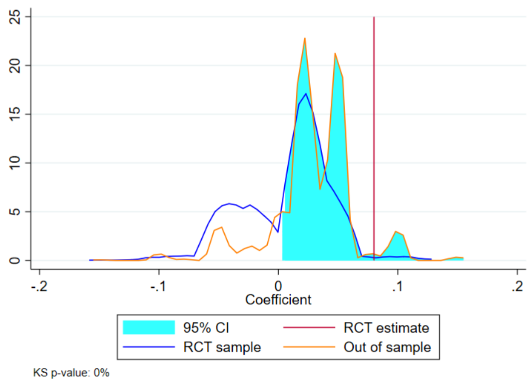
[21]The model yielded 998 valid responses for the "RCT prompt" and 1,000 valid responses for the "Out of sample prompt"

[22]We assess whether the simulated distributions are different from each other by running Kolmogorov–Smirnov (KS) tests and consistently confirm that the two distributions are statistically different. These tests confirm that the model is able to generate a different set of predictions according to the nature of the reviewer.

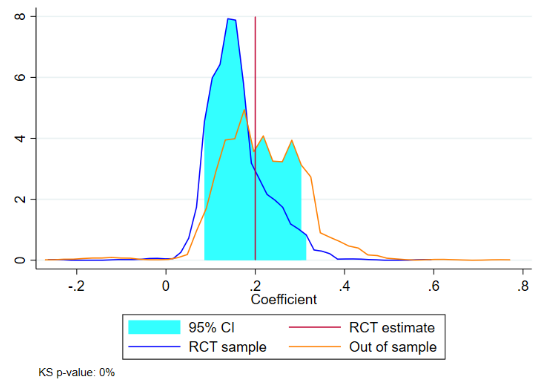
GPT's predictions. Furthermore, we show in Figure A.3 in Appendix A.5.3 that the model consistently predicts values for β_1 to be statistically insignificant while simulated results for β_2 are found to be statistically different from zero, closely matching the significance levels we report in Table 1.

Figure 7: LLM simulated coefficients

Panel A: Overall score

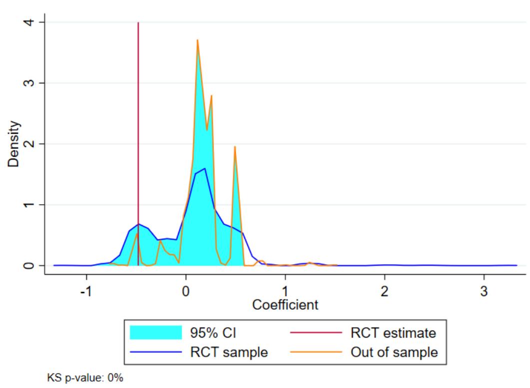


(a) Simulated β_1

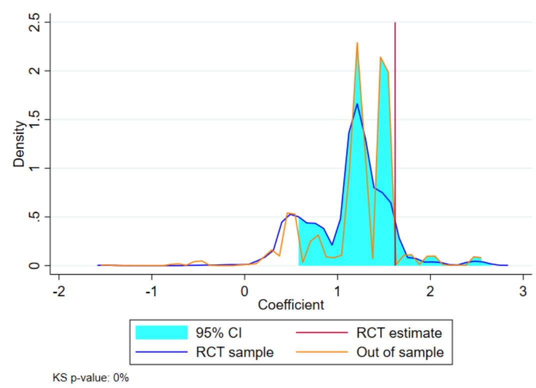


(b) Simulated β_2

Panel A: Suggestion for inclusion



(c) Simulated β_1



(d) Simulated β_2

Note: The graphs display kernel density distributions of the simulated β_1 and β_2 for Overall score and Suggestion for inclusion, as predicted across simulations. The blue line represents the distribution of results from the "RCT prompt," while the orange line corresponds to the "Out of Sample prompt." The vertical line indicates the estimated coefficient, and the shaded area reports the 95% confidence interval. We report p-values from Kolmogorov–Smirnov (KS) tests for equality of distributions at the bottom of each graph.

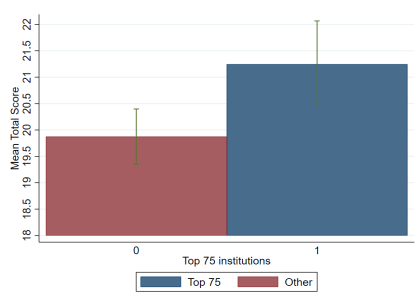
5 Mechanisms

Next, we explore the mechanisms behind the differential treatment towards top universities documented in the previous sections and characterize it. Over the many years in which the study of biases and discrimination has interested economists, two main models have been proposed to explain them: the statistical discrimination theory (Phelps, 1972; Arrow, 1973), and the taste-based discrimination theory (Becker, 1957).

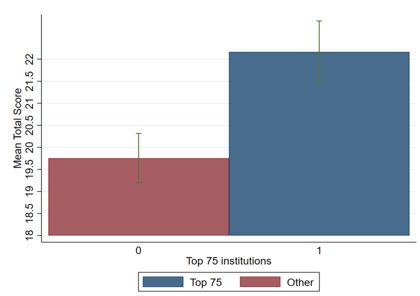
Statistical discrimination arises when beliefs on the performance of two groups diverge. In our case, if reviewers believe that papers from top-ranked institutions are better on average, they might use that belief to rate papers differently depending on the institution they observe when grading. Taste-based discrimination, on the other hand, appears when the differences in assessment between two groups are not explained by beliefs but by reviewers' preferences. In this case, just knowing the institutional affiliation of the submitter would be enough for affiliation bias to manifest.

We test for the existence of actual differences in research quality that could drive statistical discrimination by comparing the NVA reviews of papers from top and non-top institutions. Figure 8 compares the gap in scores between papers from top vs non-top universities for the VA and the NVA groups. We find that differences in evaluations, although magnified in the visible affiliation group, are already present and statistically significant when affiliation information is not available to the grader. This pattern is consistent with the presence of statistical discrimination which would arise if reviewers internalized this belief and used it in their grading.

Figure 8: Gap between Overall scores from top and non-top papers, by treatment arm



(a) Non-Visible Affiliation



(b) Visible affiliation

Note: Figure reports the average Overall score awarded to papers, divided by the treatment group of the reviewer, and by whether the author comes from a Top-75 university according to the QS Subject ranking. 95% confidence intervals for the means are included.

Statistical discrimination theory predicts that the availability of information on the objective quality of papers reduces the scope for bias, as individuals would need to rely less on the signal conveyed by the university of affiliation. Following this reasoning, we test whether our estimates of affiliation bias are driven by scarcity of information by studying the role of paper length and similarity between the paper and the reviewer's own work. We compute the word count for each paper, and the cosine similarity between the paper under review and the text submitted by the reviewer. We then divide the full sample by the median length and similarity respectively, and repeat our analysis in both sub-samples focusing on Overall score.[23]

The results, in Figure 9 below, show that affiliation bias is statistically indistinguishable in both sub-samples. This finding indicates that having a longer text to evaluate or being familiar with the paper's topics and/or methodology does not reduce the differential treatment towards top universities. In our context, in which reviewers are asked to assess extended abstracts, a possibility is that the coefficient might be capturing the informational value of the affiliation signal. The limited extension of papers prevents detailed explanation of topical and methodological issues. In addition, the lack of familiarity with the related literature can increase the salience of the signal. Statistical discrimination would predict affiliation bias to be stronger in cases where submissions are shorter or more distant to the reviewer's expertise. The lack of heterogeneity in the results is consistent with the presence of additional channels besides statistical discrimination. In addition to this, we interpret the positive and significant effect we find for the writing clarity score (Table 2 in Section 4) as support for the interpretation of β_2 as not entirely driven by statistical discrimination. Writing clarity is an aspect of the paper that should be fully inferable from the available text regardless of length or familiarity with the topic and/or methodology. In this case there is no scope for affiliation, or any other signal, to influence the grading process through statistical discrimination, strengthening the claim for additional mechanisms at play such as taste discrimination. Further explorations to fully disentangle statistical and taste discrimination are not possible in our setting, due to the limited sample size, and to the fact that we do not collect specific information on beliefs on the distribution of quality across both types of institutions, nor direct preferences for certain universities.

We test for the explicit or implicit nature of the bias by leveraging a set of questions about the importance of different elements that can be observed or inferred from the first page of a paper[24] on forming an assessment of research quality. We explore heterogeneous results according to the importance reviewers pay to affiliation when evaluating research.[25] We find that affiliation bias is only statistically distinguishable from zero for reviewers who consider affiliation important, suggesting that the participants who drive the results seem to be aware they use affiliation to form judgments on the quality of research.

[23] Position in the ranking does not only depend on the evaluation of the individual paper, but is affected by the composition of the pool of papers, and hence complicates the interpretation of results when the sample is split.

[24] We report descriptive statistics for the importance of title, number of authors, gender, nationality, and seniority of author(s), institutional affiliation of author(s), journal and acknowledgments when assessing research quality in Tables A.1 - A.3 in the Appendix. The scale of answers ranges from 1 "Not important at all" to 5 "Extremely important".

[25] We classify reviewers as *Low importance* if they rated affiliation between 1 and 2 and *High importance* if they gave a rating between 4 and 5

Figure 9: Heterogeneity by paper length and similarity with the reviewers' research

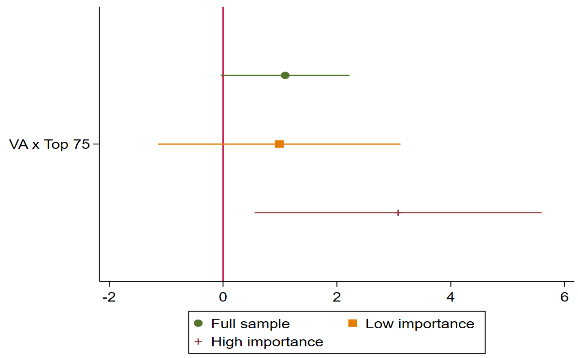
(a) By paper lenght (b) By paper-reviewer similarity

Note: Figures report 95% confidence intervals for estimates of the β_2 coefficient of our main equation, Equation 1, both for the full sample, and for the sub-samples above and below the median of the relevant dimension. The outcome is Overall score as defined in Table 1. *Short papers* are papers with a word count equal or lower than 1367. *Poor matches* are reviews in which the paper under consideration and the text submitted by the reviewer have a cosine similarity equal or lower than 0.034. As in Section 4.3, the paper-by-block fixed effects have to be substituted by paper fixed effects. We refer the reader to that section for an explanation.

Next, we explore whether our estimates reflect a strategic behavior by the reviewer, who is also a potential conference participant and may be interested in meeting and networking with peers from high-ranked institutions. To test this hypothesis, we leverage the question "Would you be interested in discussing about research with the author of this paper?". We show that, while selective reviewers drive the differential treatment reported in the results section, the ones that manifest a preferential interest to discuss research with highly-ranked peers are from the non-selective group. The evidence we provide in this section suggests the results are not driven by a stronger preference for networking with highly-ranked peers. Our findings do not imply such a preference does not exist among selective reviewers, but rather that our estimates are not directly driven by it. Selective reviewers may not have an incentive to deliberately favor applicants from top universities in the grading process of this conference if they already have professional connections to researchers from similar institutions.

Put together, these results suggest that our estimates are compatible with statistical discrimination and with suggestive evidence of other types of discrimination. In particular, the presence of a sizable effect also when reviewers and paper are close in similarity, coupled with the result for the writing score we discussed in Section 4, suggests that parts of affiliation bias can be ascribed to a taste for research from top institutions. In general, this differential treatment towards highly-ranked universities reflects an explicit bias, that is nevertheless not driven by self-interest or strategic behavior for networking.

Figure 10: Heterogeneity by reviewers' self-declared reliance on affiliation to form judgments about the quality of research.



Note: Figures report 95% confidence intervals for estimates of the β_2 coefficient of our main equation, Equation 1. The outcome is Overall score as defined in Table 1. We asked reviewers to grade from 1 "Not important at all" to 5 "Extremely important" how important affiliation is, in addition to the technical content, when reading a paper to formulate their assessment. We classify reviewers as *Low importance* if they rated affiliation between 1 and 2 and *High importance* if they gave a rating between 4 and 5. As in Section 4.3, the paper-by-block fixed effects have to be substituted by paper fixed effects. We refer the reader to that section for an explanation.

Table 4: Willing to meet with the author

	(1) Meet the Author	(2) Meet the Author	(3) Meet the Author
VA	-0.06	0.08	-0.12*
	(0.05)	(0.06)	(0.06)
VA × *Top 75*	0.16**	-0.06	0.30***
	(0.06)	(0.12)	(0.07)
Reviewer type	All	Selective	Non selective
N	1107	400	701
Control mean	0.51	0.44	0.56

Note: Dependent variable for to a dummy equal 1 for reviewers replying "Yes" to the question *"Would you be interested in discussing about research with the author of this submission?"*. The control mean refers to the average outcome in NVA grading. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin, a stratification dummy for selective reviewers and paper FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

6 Robustness checks

Our main results remain similar across multiple robustness checks. Specifically, our point estimates are virtually unaffected, although less precisely estimated, when removing reviewer controls or replacing paper-by-block fixed effects with paper fixed effects, as shown in Table A.7 in the Appendix. Following de Chaisemartin and Ramírez-Cuéllar (2024), we report p-values from clustering our standard errors at the couple level whenever we drop block fixed effects and only exploit variation within papers across all reviewers. This alternative clustering structure does not change the statistical significance of our results. Likewise, our results remain statistically significant when we adjust p-values for multiple hypothesis testing as shown in Table A.8 in the Appendix.

Additionally, we replicate our procedure for selecting the cutoff for elite institutions by using narrower bins of 15 positions in the ranking, as shown in Figure **??** in the Appendix. This finer binning suggests a cutoff of 60 for selective institutions, similar to the choice of 75 used in our main specification. Figure A.4 in the Appendix demonstrates the robustness of our main results to different cutoff choices. Overall, our point estimates remain stable when using more stringent definitions for elite institutions, although the estimated confidence intervals grow larger, reflecting a reduction in power due to the smaller number of papers from those institutions.

We next address possible concerns on violation of treatment allocation, increased signal salience and experimenter demand effect. We start by dropping the 20 papers that could be found online at the end of the grading phase, and hence potentially traceable to their authors. The low percentage of works available online, and the stability of the estimates when we drop these observations, confirm that our exercise is well suited to ensure the compliance of the treatment allocation. Next, we tackle concerns on having increased the salience of affiliation as an information signal in our design. We drop from the sample reviewers who mentioned affiliation in the open-ended section of the post-grading survey and replicate our main specification. Our coefficient remains qualitatively stable despite growing imprecise. Similarly, to ensure that results are not driven by experimenter demand effect (Mummolo and Peterson, 2019; Zizzo, 2010), we replicate our main specification after dropping from the sample respondents who made reference to the possibility of an experiment in the open-ended fields of the post-grading survey. Experimenter demand effect should not be a threat to our estimates, given that participants were unaware of the experiment. Nevertheless, if a large fraction of reviewers were suspicious about being part of an experiment it may affect how they evaluated papers and bias our estimates. Overall, even with this conservative approach, our estimates remain qualitatively similar, with a slight loss of precision due to the reduced sample size. Hence, we conclude that experimenter demand does not affect our findings. The results of these exercises can be found in Figure A.5 in the Appendix.

Finally, we turn to the rankings we used to define papers from top institutions. First, we replicate our main estimates across the four sub-indicators of the QS subject specific score:

“Academic reputation”, “Employer reputation”, “Citations”, and “H-index”.[26] Across all sub-indicators we find sizable effects for affiliation bias as shown in Table A.6.1 in the Appendix, suggesting that a single sub-indicator does not drive our overall effect. Next, we repeat our analysis using six well-known alternative university rankings to define our “Top institutions” category. Despite differences in the methodologies and the criteria used, results largely hold.[27]

7 Conclusion

Conferences are crucial events in a scholar’s career, serving as platforms for networking, exchange of ideas, and visibility. Admission to conferences typically relies on peer review, a process susceptible to bias and misperception, which can profoundly impact individual careers and the production of scientific knowledge. Existing research shows the extent to which labels and stereotypes, such as gender (Hóspido and Sanz, 2020; Samahita and Devereux, 2024; Card et al., 2020, 2022), seniority (Seeber and Bacchelli, 2017; Uchida, 2021), nationality (Tavoletti et al., 2022), ethnicity (Pleskac et al., 2024), physical appearance (Hale et al., 2023), and author’s prominence (Huber et al., 2022; Tomkings et al., 2017), can influence the evaluation of research in the peer review process and may contribute to the lack of diversity in academia.

In this paper, we focus on affiliation, a salient label in economics with strong potential for bias, and whose causal impact is currently understudied. We also take a closer look at conference inclusions, another important aspect of the academic profession which has been mostly overlooked from a research standpoint. Existing studies have mainly focused on the effects of biases on publications in top journals and citations (Blank, 1991; Card et al., 2020, 2022; Ersoy and Pate, 2023), while evidence of how these biases could affect admission to conferences is less developed, and restricted to disciplines other than Economics (Uchida, 2021) or different types of biases (Chari and Goldsmith-Pinkham, 2017; Pleskac et al., 2024). To the best of our knowledge, our study is among the few papers that produce robust causal evidence of the effect of displaying affiliation on the evaluation of research. Moreover, our findings are the first to show that affiliation bias is not exclusive of senior, established scholars, but a pervasive behavior that starts at the doctoral stage. Our evidence documents how affiliation bias has profound implications for conferences’ composition and the perpetuation of inequalities in academia.

In this paper, we present the results of a field experiment designed to detect and quantify the presence of affiliation bias in the peer review process of an early-career workshop in Economics at the University Paris 1 Pantheon-Sorbonne. We implemented a matched-pair

[26] We point the reader to a description of the methodology of the QS Subject-specific ranking in Appendix A.6.1.

[27] Appendix A.6.2 provides a comparison of “selective” universities in our sample across the different rankings

experimental design by creating couples of reviewers, consisting of a reviewer assigned to visible affiliation (VA) grading and another assigned to Non-Visible Affiliation (NVA) grading. We removed any further identifying information in all papers. Each couple received a block of eight randomly allocated papers from the pool of applications to the workshop. This design allows us to identify the causal effect of visible affiliation grading by exploiting variation within papers and allowing at the same time for grading on a curve.

We find evidence that papers from prestigious institutions get higher grades when they display their affiliation. In particular, papers from Top 75 institutions receive an average additional premium in their Overall score of 1.62 points on a 30-point scale, moving up 18.59 (12%) positions in the ranking, compared to papers from less prestigious institutions. These effects ultimately shape conference admissions. In a simulation exercise where we apply to our sample acceptance rates from the annual meetings of the American Economic Association and the European Economic Association, affiliation bias improves the probability of conference inclusion for applicants from top institutions by 12.9 and 18.2 percentage points respectively. This bias in turn has important consequences for the diversity of conference participants, notably by curtailing the participation of women and first-generation attendees, which are already underrepresented in Economics and particularly so in Top universities (Schultz and Stansbury, 2022; Ersoy and Pate, 2023).

Next, we investigate the mechanisms underpinning our results. We find that papers from selective institutions score higher on blind grading assessment which we take as an indication of higher objective research quality. We argue that reviewers make use of this generalization and grade papers differently based on this result, hinting at a role for statistical discrimination. We further investigate this result by studying heterogeneous effects across the similarity of reviewers' own work and the paper they are asked to evaluate. Statistical discrimination theory would predict that the role of affiliation bias should widen the less information reviewers have, or equivalently, the more distant they are from the paper they are asked to evaluate. However, we find availability of information to play a limited role, which suggests that there might be room for additional types of discrimination such as taste-based. We then show that affiliation bias is stronger among reviewers who declare affiliation to be an important signal when evaluating research, but we rule out any strategic behavior coming from the willingness to network with other participants. Finally, when we split the sample between selective and non-selective reviewers, we find that selective reviewers drive the bias, pointing towards evidence of in-group bias or club-like favoritism.

Taken together, our findings highlight how affiliation bias profoundly affects conference composition in favor of papers from prestigious institutions, ultimately reducing diversity in academia and scientific knowledge. Our findings are particularly noteworthy in Economics, the discipline we study, often described as plagued by strong inequalities, elitism, and underrepresentation of minorities (Fourcade et al., 2015; Freeman et al., 2024; Hoover and Svorenčík, 2023; Schultz and Stansbury, 2022).

We argue that conferences should move to a double-blind evaluation standard, and reviewers should be informed about the cost they impose on the careers of scholars when leveraging signals to formulate their decisions. Blind evaluation is likely more effective in conferences compared to journal submissions. We believe this to be particularly true for events like ours, targeted at preliminary work by early-career researchers which, unlike papers sent for publication that can be traced back to the authors via working papers and seminars, is less likely to be publicly available. In addition, this specific subpopulation of early-career researchers is likely to benefit the most from a fair selection process across all conferences given the prominent role conferences play at this stage of academic careers (Leite Lopez de Leon and McQuillin, 2020).

Our study has some limitations that future research should explore. First, our results are conditioned by our setting and our pool of applicants, limiting their external validity. Because composition effects depend on the applicant pools, and applicants are likely to behave differently when facing well-known conferences with established acceptance rates, our simulation exercise might not fully represent what would happen across acceptance rates in a different context. Nevertheless, our findings resonate with previous ones (Blank, 1991; Ersoy and Pate, 2023; Uchida, 2021) and align closely with predictions from a large language model, alleviating concerns about their generalizability to other contexts and mirroring DellaVigna et al. (2020)'s findings on experimental results stability.

Second, without additional data, it is difficult to evaluate whether Non-Visible Affiliation or visible affiliation grading produced more accurate judgments of research potential. Related research (Uchida, 2021; Pleskac et al., 2024) uses publication and citation data to assess paper quality, acknowledging that this approach is not without limitations, as both metrics are susceptible to similar biases and do not offer an entirely objective evaluation (Card et al., 2020). Nonetheless, their findings indicate that blind review does not impair reviewers' ability to select high-quality papers, suggesting that knowledge of the author's identity does not necessarily enhance the peer-review process. In this paper, we exploit grades given under Non-Visible Affiliation grading and a measure of predicted citations derived from submissions' titles via a Machine Learning algorithm (Iaria et al., 2024; Schwarz, 2023) as a proxy for submission quality. Our findings are consistent with those found by the literature and suggest that institutional affiliation information does not improve the peer-review process.

Third, while we provide suggestive evidence linking discrimination to different explanations, our experimental setup does not allow us to formally test different theories for discrimination, an area that is receiving fresh research interest (Bordalo et al., 2016, 2019; Bohren et al., 2022; Amer et al., 2024; Barron et al., 2024; Lepage, 2024; Bartoš et al., 2016). Future research should investigate further the mechanisms behind affiliation bias in order to improve our understanding of its origin and better inform policy.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge (2023) "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics*, 138 (1), 1–35.

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai (2023) "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," `https://arxiv.org/abs/2208.10264`.

Akerlof, George A. and Rachel E. Kranton (2000) "Economics and Identity," *The Quarterly Journal of Economics*, 115 (3), 715–753, `http://www.jstor.org/stable/2586894`.

Amer, Abdelrahman, Ashley C. Craig, and Clementine Van Effenterre (2024) "Decoding Gender Bias: The Role of Personal Interaction," *IZA Discussion Paper Series No. 17077*.

Arrow, J., Kenneth (1973) *Discrimination in Labor Markets*, Chap. The Theory of Discrimination: Princeton University Press.

Barron, Kai, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch (2024) "Explicit and implicit belief-based gender discrimination: A hiring experiment," *Management Science*.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka (2016) "Attention discrimination: Theory and field experiments with monitoring information acquisition," *American Economic Review*, 106 (6), 1437–1475.

Bayer, Amanda and Cecilia Elena Rouse (2016) "Diversity in the economics profession: A new attack on an old problem," *Journal of Economic Perspectives*, 30 (4), 221–242.

Becker, Gary S. (1957) *The Economics of Discrimination*: University of Chicago.

Bellemare, Marc F (2022) *Doing economics: What you should have learned in grad school—but didn't*: MIT Press.

Bethmann, Dirk, Felix Bransch, Michael Kvasnicka, and Abdolkarim Sadrieh (2023) "Home Bias in Top Economics Journals," *IZA Discussion Paper Series, No. 15965*.

Blank, Rebecca M. (1991) "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review," *The American Economic Review*, 81 (5), 1041–1067, `https://www.jstor.org/stable/2006906`, Publisher: American Economic Association.

Bohren, J Aislinn, Peter Hull, and Alex Imas (2022) "Systemic discrimination: Theory and measurement,"Technical report, National Bureau of Economic Research.

Bonferroni, Carlo E (1935) "Il calcolo delle assicurazioni su gruppi di teste," *Studi in onore del Professore Salvatore Ortu Carboni*, 13–60.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016) "Stereotypes," *The Quarterly Journal of Economics*, 131 (4), 1753–1794.

——— (2019) "Beliefs about gender," *American Economic Review*, 109 (3), 739–773.

Calsamiglia, Caterina and Annalisa Loviglio (2019) "Grading on a curve: When having good peers is not good," *Economics of Education Review*, 73, 101916.

Calvo-Armengol, Antoni and Matthew O Jackson (2004) "Social networks in determining employment: Patterns, dynamics, and inequality," *American Economic Review*, 94 (3), 426–454.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri (2020) "Are referees and editors in economics gender neutral?" *The Quarterly Journal of Economics*, 135 (1), 269–327, `https://academic.oup.com/qje/article-abstract/135/1/269/5614978`, Publisher: Oxford University Press.

——— (2022) "Gender Differences in Peer Recognition by Economists," *Econometrica*, 90 (5), 1937–1971, 10.3982/ECTA18027.

Carrell, Scott, David Figlio, and Lester Lusher (2024) "Clubs and Networks in Economics Reviewing," *Journal of Political Economy*, 132 (7), 777–797, 10.1086/730208.

de Chaisemartin, Clément and Jaime Ramírez-Cuéllar (2024) "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics*, 16 (1), 193–212, 10.1257/app.20210252.

Chari, Anusha and Paul Goldsmith-Pinkham (2017) "Gender Representation in Economics Across Topics and Time: Evidence from the NBER Summer Institute," *NBER Working Paper No. 23953*.

Chen, Yan and Sherry Xin Li (2009) "Group Identity and Social Preferences," *American Economic Review*, 99 (1), 431–57, 10.1257/aer.99.1.431.

Chen, Yaoyu, Yuheng Hu, and Yingda Lu (2024) "Simulating Field Experiments with Large Language Models," `https://arxiv.org/abs/2408.09682`.

Chetty, Raj, Matthew O Jackson, Theresa Kuchler et al. (2022) "Social capital I: measurement and associations with economic mobility," *Nature*, 608 (7921), 108–121.

Coffman, Katherine B, Manuela R Collis, and Leena Kulkarni (2024) "Whether to apply," *Management Science*, 70 (7), 4649–4669.

Cullen, Zoe and Ricardo Perez-Truglia (2023) "The Old Boys' Club: Schmoozing and the Gender Gap," *American Economic Review*, 113, 1703–1740, 10.12578/aer.20210863.

DellaVigna, Stefano, Nicholas Otis, and Eva Vivalt (2020) "Forecasting the results of experiments: Piloting an elicitation strategy," in *AEA Papers and Proceedings*, 110, 75–79, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt (2019) "Predict science to improve science," *Science*, 366 (6464), 428–429.

Ductor, Lorenzo, Marcel Fafchamps, Sanjeev Goyal, and Marco J. van der Leij (2014) "Social Networks and Research Output," *The Review of Economics and Statistics*, 96 (5), 936–948, `https://www.jstor.org/stable/43554968`, Publisher: The MIT Press.

Ersoy, Fulya Y. and Jennifer Pate (2023) "Invisible hurdles: Gender and institutional differences in the evaluation of economics papers," *Economic Inquiry*, 61 (4), 777–797, 10.1111/ecin.13145, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.13145.

Farré, Lídia and Francesc Ortega (2024) "Geographic mobility of college students and the gender gap in academic aspirations," *Labour Economics*, 90, 102550, https://doi.org/10.1016/j.labeco.2024.102550.

Flory, Jeffrey A, Andreas Leibbrandt, and John A List (2015) "Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions," *The Review of Economic Studies*, 82 (1 (290)), 122–155, `http://www.jstor.org/stable/43551466`.

Fourcade, Marion, Etienne Ollion, and Yann Algan (2015) "The superiority of economists," *The Journal of Economic Perspectives*, 29, 89 – 114, 10.1257/jep.29.1.89.

Freeman, Richard B, Danxia Xie, Hanzhe Zhang, and Hanzhang Zhou (2024) "High and Rising Institutional Concentration of Award-Winning Economists."

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (2021) "Conference presentations and academic publishing," *Economic Modelling*, 95, 228–254, 10.1016/j.econmod.2020.12.017.

Hale, Galina, Tali Regev, and Yona Rubinstein (2023) "Do looks matter for an academic career in economics?" *Journal of Economic Behavior & Organization*, 215, 406–420, 10.1016/j.jebo.2023.09.022.

Head, Keith, Yao Amber Li, and Asier Minondo (2019) "Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics," *The Review of Economics and Statistics*, 101 (4), 713–727.

Holm, Sture (1979) "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 65–70.

Hoover, Kevin D and Andrej Svorenčík (2023) "Who runs the AEA?" *Journal of Economic Literature*, 61 (3), 1127–1171.

Horton, John J (2023) "Large language models as simulated economic agents: What can we learn from homo silicus?"Technical report, National Bureau of Economic Research.

Hóspido, Laura and Carlos Sanz (2020) "Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences," *Oxford Bulletin of Economics and Statistics*, 83, 590–618.

Huber, Jürgen, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith (2022) "Nobel and novice: Author prominence affects peer review," *Proceedings of the National Academy of Sciences*, 119 (41), e2205779119, 10.1073/pnas.2205779119, Publisher: Proceedings of the National Academy of Sciences.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger (2024) "Gender gaps in academia: Global evidence over the twentieth century," *Available at SSRN 4150221*.

Lawson, Nicholas (2023) "What citation tests really tell us about bias in academic publishing," *European Economic Review*, 158, 104534.

Leite Lopez de Leon, Fernanda and Ben McQuillin (2020) "The Role of Conferences on the Pathway to Academic Impact," *The Journal of Human Resources*, 55 (1), 164 – 193, https://doi.org/10.3368/jhr.55.1.1116-8387R.

Lepage, Louis P. (2024) "Experience-based discrimination," *American Economic Journal: Applied Economics*, 16 (4), 288–321.

Lippert, Steffen, Anna Dreber, Magnus Johannesson, Warren Tierney, Wilson Cyrus-Lai, Eric Luis Uhlmann, Emotion Expression Collaboration, and Thomas Pfeiffer (2024) "Can large language models help predict results from a complex behavioural science study?" *Royal Society Open Science*, 11 (9), 240682.

Lleras-Muney, Adriana, Matthew Miller Miller, Shuyang Sheng, and Veronica T. Sovero (2020) "Party On: The Labor Market Returns to Social Networks and Socializing," *NBER Working Paper No. 27337*.

Mummolo, Jonathan and Erik Peterson (2019) "Demand Effects in Survey Experiments: An Empirical Assessment," *American Political Science Review*, 113 (2), 517–529, 10.1017/S0003055418000837.

Phelps, Edmund S. (1972) "The statistical theory of racism and sexism," *American Economic Review*, 62, 659–661, 10.12578/aer.20171339.

Pleskac, Timothy J, Ellie Kyung, Gretchen Chapman, and Oleg Urminsky (2024) "Blinded versus unblinded review: A field study comparing the equity of peer-review," *University of Chicago, Becker Friedman Institute for Economics Working Paper No.2024-07.*

Ray, Debraj ⓡ Arthur Robson (2018) "Certified Random: A New Order for Coauthorship," *American Economic Review*, 108 (2), 489–520, 10.1257/aer.20161492.

Reingewertz, Yaniv and Carmela Lutmar (2018) "Academic in-group bias: An empirical examination of the link between author and journal affiliation," *Journal of Informetrics*, 12 (1), 74–86, 10.1016/j.joi.2017.11.006.

Rose, Michael E. and Suraj Shekhar (2023) "Adviser connectedness and placement outcomes in the economics job market," *Labour Economics*, 84.

Samahita, Margaret and Kevin Devereux (2024) "Are Economics Conferences Gender-Neutral? Evidence from Ireland*," *Oxford Bulletin of Economics and Statistics*, 86 (1), 101–118, 10.1111/obes.12575, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12575.

Schultz, Robert and Anna Stansbury (2022) "Socioeconomic diversity of economics PhDs," *Peterson Institute for International Economics Working Paper* (22-4).

Schwarz, Carlo (2023) "Estimating text regressions using txtreg_train," *The Stata Journal*, 23 (3), 799–812.

Seeber, Marco and Alberto Bacchelli (2017) "Does single blind peer review hinder newcomers?" *Scientometrics*, 113 (1), 567–585, 10.1007/s11192-017-2264-7.

Spence, Michael (1973) "Job Market Signaling," *The Quarterly Journal of Economics*, 87 (3), 355–374.

Tavoletti, Ernesto, Robert D. Stephens, Vas Taras, and Longzhu Dong (2022) "Nationality biases in peer evaluations: The country-of-origin effect in global virtual teams," *International Business Review*, 31 (2), 101969, 10.1016/j.ibusrev.2021.101969.

Tomkings, Andrew, Min Zhang, and William D. Heavlin (2017) "Reviewer bias in single-versus double-blind peer review," *Proceedings of the National Academy of Sciences USA*, 114, 12708–12713, 10.1073/pnas.1707323114.

Tsugawa, Sho, Takuya Kanetsuki, and Junichi Sugihara (2022) "Relationship between early-career collaboration among researchers and future funding success in Japanese academia," *PloS One*, 17 (11), e0277621, 10.1371/journal.pone.0277621.

Uchida, Haruka (2021) “What Do Names Reveal? Impacts of Blind Evaluations on Composition and Quality,” January, 10.2139/ssrn.3767565.

Zinovyeva, Natalia and Manuel Bagues (2015) “The Role of Connections in Academic Promotions,” *American Economic Journal: Applied Economics*, 7 (2), 264–92, 10.1257/app.20120337.

Zizzo, Daniel John (2010) “Experimenter demand effects in economic experiments,” *Experimental Economics*, 13, 75–98, 10.1007/s10683-009-9230-z.

A Appendix

A.1 Descriptive statistics

Table A.1: Descriptive Statistics by Treatment Status

	All		Visible affiliation		Non-visible affiliation		Difference in means	
Variable	N	Mean	N	Mean	N	Mean	Difference	P-value
University ranked in QS	140	0.76	68	0.72	72	0.81	-0.09	0.24
Position of University in QS Ranking	107	172.58	49	192.02	58	156.16	35.87	0.22
Gender: Male	140	0.72	68	0.72	72	0.72	-0.00	0.98
Gender: Female	140	0.28	68	0.28	72	0.28	0.00	0.98
Parents ED: Not tertiary	140	0.37	68	0.35	72	0.39	-0.04	0.66
Parents ED: Tertiary	140	0.62	68	0.63	72	0.61	0.02	0.80
Parents ED: Not reported	140	0.01	68	0.01	72	0.00	0.01	0.31
Country: Not reported	140	0.06	68	0.07	72	0.04	0.03	0.42
Country: Eastern Europe	140	0.06	68	0.09	72	0.03	0.06	0.13
Country: Northern Europe	140	0.04	68	0.03	72	0.04	-0.01	0.70
Country: Outside of Europe	140	0.36	68	0.40	72	0.32	0.08	0.34
Country: Southern Europe	140	0.30	68	0.22	72	0.38	-0.15	0.05
Country: Western Europe	140	0.19	68	0.19	72	0.19	-0.00	0.96
PhD start year: 2014-2018	139	0.18	68	0.19	71	0.17	0.02	0.74
PhD start year: 2019	139	0.22	68	0.24	71	0.20	0.04	0.59
PhD start year: 2020	139	0.19	68	0.18	71	0.21	-0.03	0.61
PhD start year: 2021	139	0.22	68	0.24	71	0.20	0.04	0.59
PhD start year: 2022-2023	139	0.19	68	0.15	71	0.23	-0.08	0.24
PhD start year: Not reported	139	0.01	68	0.01	71	0.00	0.01	0.31
Survey response	156	0.90	78	0.87	77	0.94	-0.06	0.19
Liked Peer Review	140	7.53	68	7.50	72	7.56	-0.06	0.88
Liked Censoring	140	8.91	68	8.76	72	9.06	-0.29	0.33
Liked Conference Open to Phd and Post Docs	140	9.31	68	9.28	72	9.35	-0.07	0.77
Liked Submission of Extended Abstracts	140	7.78	68	7.81	72	7.75	0.06	0.88
Imp. Affiliation for the Reviewer	140	2.66	68	2.62	72	2.69	-0.08	0.71
Imp. Journal for the Reviewer	140	3.66	68	3.65	72	3.67	-0.02	0.92
Imp. the Number of Authors for the Reviewer	140	1.97	68	2.07	72	1.88	0.20	0.27
Imp. Ackowledgments for the Reviewer	140	1.80	68	1.72	72	1.88	-0.15	0.39
Imp. Title for the Reviewer	140	3.19	68	3.09	72	3.29	-0.20	0.36
Imp. Gender for the Reviewer	140	1.18	68	1.16	72	1.19	-0.03	0.75
Imp. Nationality for the Reviewer	140	1.21	68	1.24	72	1.19	0.04	0.63
Imp. Seniority for the Reviewer	140	2.25	68	2.22	72	2.28	-0.06	0.77
Imp. Affiliation in Economics	139	4.14	67	4.24	72	4.04	0.20	0.29
Imp. Journal in Economics	139	4.54	67	4.54	72	4.54	-0.00	0.98
Imp. the Number of Authors in Economics	139	2.80	67	2.76	72	2.83	-0.07	0.72
Imp. Ackowledgments in Economics	139	2.48	67	2.37	72	2.58	-0.21	0.30
Imp. Title in Economics	139	3.23	67	3.25	72	3.21	0.05	0.82
Imp. Gender in Economics	139	2.27	67	2.22	72	2.31	-0.08	0.70
Imp. Nationality in Economics	139	2.42	67	2.60	72	2.26	0.33	0.13
Imp. Seniority in Economics	139	3.88	67	3.88	72	3.88	0.01	0.98
Imp. Joining Conferences	139	4.34	67	4.30	72	4.38	-0.08	0.58
Imp. Building a Network	139	4.56	67	4.46	72	4.65	-0.19	0.07
Imp. Supervisor's Network	139	4.60	67	4.55	72	4.65	-0.10	0.38
Number of Conferences One Would Like to Attend	139	3.73	67	3.82	72	3.65	0.17	0.52
Willingness to Meet an Econmist They Don't Know	139	4.62	67	4.61	72	4.63	-0.01	0.92
Willingness to Have Dinner with a Visiting Scholar	139	4.35	67	4.30	72	4.39	-0.09	0.57

Note: Table reports descriptive statistics at the reviewer level broken down by treatment arm. It also includes mean differences and the corresponding p-values. The table highlights reviewers' demographic characteristics such as gender, parental education, and country of origin. We also gathered information on reviewers' preferences, rated on a scale from 1 to 10, regarding the following aspects of the conference submission process: peer review, anonymization of submissions, the inclusivity of PhD students and Postdocs, and the submission of extended abstracts. Additionally, reviewers rated the importance of several signals for themselves and the economics profession on a scale from 1 (Not important at all) to 5 (Extremely important). Finally, the table includes information on the number of conferences reviewers wish to attend and their willingness to network (rated from 1 (Not at all) to 5 (Very much)) through meetings or dinners with other economists or visiting scholars.

Table A.2: Descriptive Statistics by Selectivity of Reviewers

	All		Selective		Non-selective		Difference in means	
Variable	N	Mean	N	Mean	N	Mean	Difference	P-value
University ranked in QS	140	0.76	51	0.94	89	0.66	0.28	0.00
Position of University in QS Ranking	107	172.58	48	47.94	59	273.98	-226.05	0.00
Gender: Male	140	0.72	51	0.78	89	0.69	0.10	0.21
Gender: Female	140	0.28	51	0.22	89	0.31	-0.10	0.21
Parents ED: Not tertiary	140	0.37	51	0.18	89	0.48	-0.31	0.00
Parents ED: Tertiary	140	0.62	51	0.82	89	0.51	0.32	0.00
Parents ED: Not reported	140	0.01	51	0.00	89	0.01	-0.01	0.45
Country: Not reported	140	0.06	51	0.04	89	0.07	-0.03	0.49
Country: Eastern Europe	140	0.06	51	0.04	89	0.07	-0.03	0.49
Country: Northern Europe	140	0.04	51	0.06	89	0.02	0.04	0.27
Country: Outside of Europe	140	0.36	51	0.37	89	0.35	0.02	0.78
Country: Southern Europe	140	0.30	51	0.24	89	0.34	-0.10	0.21
Country: Western Europe	140	0.19	51	0.25	89	0.16	0.10	0.16
PhD start year: 2014-2018	139	0.18	51	0.20	88	0.17	0.03	0.71
PhD start year: 2019	139	0.22	51	0.16	88	0.25	-0.09	0.20
PhD start year: 2020	139	0.19	51	0.16	88	0.22	-0.06	0.40
PhD start year: 2021	139	0.22	51	0.22	88	0.22	0.00	1.00
PhD start year: 2022-2023	139	0.19	51	0.27	88	0.14	0.14	0.04
PhD start year: Not reported	139	0.01	51	0.00	88	0.01	-0.01	0.45
Survey response	156	0.90	55	0.93	100	0.89	0.04	0.46
Liked Peer Review	140	7.53	51	7.33	89	7.64	-0.31	0.42
Liked Censoring	140	8.91	51	8.84	89	8.96	-0.11	0.72
Liked Conference Open to Phd and Post Docs	140	9.31	51	9.16	89	9.40	-0.25	0.31
Liked Submission of Extended Abstracts	140	7.78	51	7.92	89	7.70	0.22	0.58
Imp. Affiliation for the Reviewer	140	2.66	51	2.57	89	2.71	-0.14	0.51
Imp. Journal for the Reviewer	140	3.66	51	3.57	89	3.71	-0.14	0.51
Imp. the Number of Authors for the Reviewer	140	1.97	51	1.82	89	2.06	-0.23	0.21
Imp. Acknowledgments for the Reviewer	140	1.80	51	1.53	89	1.96	-0.43	0.02
Imp. Title for the Reviewer	140	3.19	51	3.25	89	3.16	0.10	0.67
Imp. Gender for the Reviewer	140	1.18	51	1.10	89	1.22	-0.13	0.23
Imp. Nationality for the Reviewer	140	1.21	51	1.14	89	1.26	-0.12	0.17
Imp. Seniority for the Reviewer	140	2.25	51	2.24	89	2.26	-0.02	0.91
Imp. Affiliation in Economics	139	4.14	51	4.25	88	4.07	0.19	0.34
Imp. Journal in Economics	139	4.54	51	4.67	88	4.47	0.20	0.20
Imp. the Number of Authors in Economics	139	2.80	51	2.92	88	2.73	0.19	0.36
Imp. Acknowledgments in Economics	139	2.48	51	2.35	88	2.56	-0.20	0.33
Imp. Title in Economics	139	3.23	51	3.00	88	3.36	-0.36	0.08
Imp. Gender in Economics	139	2.27	51	2.31	88	2.24	0.08	0.73
Imp. Nationality in Economics	139	2.42	51	2.47	88	2.40	0.07	0.75
Imp. Seniority in Economics	139	3.88	51	3.92	88	3.85	0.07	0.75
Imp. Joining Conferences	139	4.34	51	4.16	88	4.44	-0.29	0.04
Imp. Building a Network	139	4.56	51	4.55	88	4.57	-0.02	0.86
Imp. Supervisor's Network	139	4.60	51	4.69	88	4.56	0.13	0.27
Number of Conferences One Would Like to Attend	139	3.73	51	3.73	88	3.74	-0.01	0.96
Willingness to Meet an Economist They Don't Know	139	4.62	51	4.63	88	4.61	0.01	0.92
Willingness to Have Dinner with a Visiting Scholar	139	4.35	51	4.29	88	4.38	-0.08	0.63

Notes: Table reports descriptive statistics at the reviewer level broken down by selectivity of reviewer type. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. It also includes mean differences and the corresponding p-values. The table highlights reviewers' demographic characteristics such as gender, parental education, and country of origin. We also gathered information on reviewers' preferences, rated on a scale from 1 to 10, regarding the following aspects of the conference submission process: peer review, anonymization of submissions, the inclusivity of PhD students and Postdocs, and the submission of extended abstracts. Additionally, reviewers rated the importance of several signals for themselves and the economics profession on a scale from 1 (Not important at all) to 5 (Extremely important). Finally, the table includes information on the number of conferences reviewers wish to attend and their willingness to network (rated from 1 (Not at all) to 5 (Very much)) through meetings or dinners with other economists or visiting scholars.

Table A.3: Descriptive Statistics by Top 75 Universities

	All		Top 75		Non-Top 75		Difference in means	
Variable	N	Mean	N	Mean	N	Mean	Difference	P-value
University ranked in QS	140	0.76	32	1.00	108	0.69	0.31	0.00
Position of University in QS Ranking	107	172.58	32	31.16	75	232.92	-201.76	0.00
Gender: Male	140	0.72	32	0.75	108	0.71	0.04	0.68
Gender: Female	140	0.28	32	0.25	108	0.29	-0.04	0.68
Parents ED: Not tertiary	140	0.37	32	0.09	108	0.45	-0.36	0.00
Parents ED: Tertiary	140	0.62	32	0.91	108	0.54	0.37	0.00
Parents ED: Not reported	140	0.01	32	0.00	108	0.01	-0.01	0.59
Country: Not reported	140	0.06	32	0.03	108	0.06	-0.03	0.48
Country: Eastern Europe	140	0.06	32	0.00	108	0.07	-0.07	0.11
Country: Northern Europe	140	0.04	32	0.09	108	0.02	0.08	0.04
Country: Outside of Europe	140	0.36	32	0.44	108	0.33	0.10	0.28
Country: Southern Europe	140	0.30	32	0.22	108	0.32	-0.11	0.26
Country: Western Europe	140	0.19	32	0.22	108	0.19	0.03	0.68
PhD start year: 2014-2018	139	0.18	32	0.19	107	0.18	0.01	0.90
PhD start year: 2019	139	0.22	32	0.13	107	0.24	-0.12	0.16
PhD start year: 2020	139	0.19	32	0.16	107	0.21	-0.05	0.54
PhD start year: 2021	139	0.22	32	0.22	107	0.21	0.00	0.96
PhD start year: 2022-2023	139	0.19	32	0.31	107	0.15	0.16	0.04
PhD start year: Not reported	139	0.01	32	0.00	107	0.01	-0.01	0.59
Survey response	156	0.90	34	0.94	121	0.89	0.05	0.40
Liked Peer Review	140	7.53	32	7.53	108	7.53	0.00	0.99
Liked Censoring	140	8.91	32	8.75	108	8.96	-0.21	0.55
Liked Conference Open to PhD and Postdocs	140	9.31	32	9.41	108	9.29	0.12	0.67
Liked Submission of Extended Abstracts	140	7.78	32	7.56	108	7.84	-0.28	0.55
Importance of Affiliation for the Reviewer	140	2.66	32	2.63	108	2.67	-0.04	0.86
Importance of Journal for the Reviewer	140	3.66	32	3.75	108	3.63	0.12	0.62
Importance of the Number of Authors for the Reviewer	140	1.97	32	1.94	108	1.98	-0.04	0.84
Importance of Acknowledgments for the Reviewer	140	1.80	32	1.56	108	1.87	-0.31	0.14
Importance of Title for the Reviewer	140	3.19	32	3.41	108	3.13	0.28	0.30
Importance of Gender for the Reviewer	140	1.18	32	1.16	108	1.19	-0.03	0.81
Importance of Nationality for the Reviewer	140	1.21	32	1.22	108	1.21	0.01	0.95
Importance of Seniority for the Reviewer	140	2.25	32	2.53	108	2.17	0.36	0.12
Importance of Affiliation in Economics	139	4.14	32	4.50	107	4.03	0.47	0.03
Importance of Journal in Economics	139	4.54	32	4.81	107	4.46	0.35	0.04
Importance of the Number of Authors in Economics	139	2.80	32	3.19	107	2.68	0.51	0.04
Importance of Acknowledgments in Economics	139	2.48	32	2.41	107	2.50	-0.10	0.68
Importance of Title in Economics	139	3.23	32	3.00	107	3.30	-0.30	0.20
Importance of Gender in Economics	139	2.27	32	2.34	107	2.24	0.10	0.69
Importance of Nationality in Economics	139	2.42	32	2.41	107	2.43	-0.02	0.93
Importance of Seniority in Economics	139	3.88	32	4.13	107	3.80	0.32	0.20
Importance of Joining Conferences	139	4.34	32	3.97	107	4.45	-0.48	0.00
Importance of Building a Network	139	4.56	32	4.47	107	4.59	-0.12	0.33
Importance of Supervisor's Network	139	4.60	32	4.63	107	4.60	0.03	0.84
Number of Conferences One Would Like to Attend	139	3.73	32	3.75	107	3.73	0.02	0.95
Willingness to Meet an Economist They Don't Know	139	4.62	32	4.59	107	4.63	-0.03	0.83
Willingness to Have Dinner with a Visiting Scholar	139	4.35	32	4.22	107	4.38	-0.16	0.39

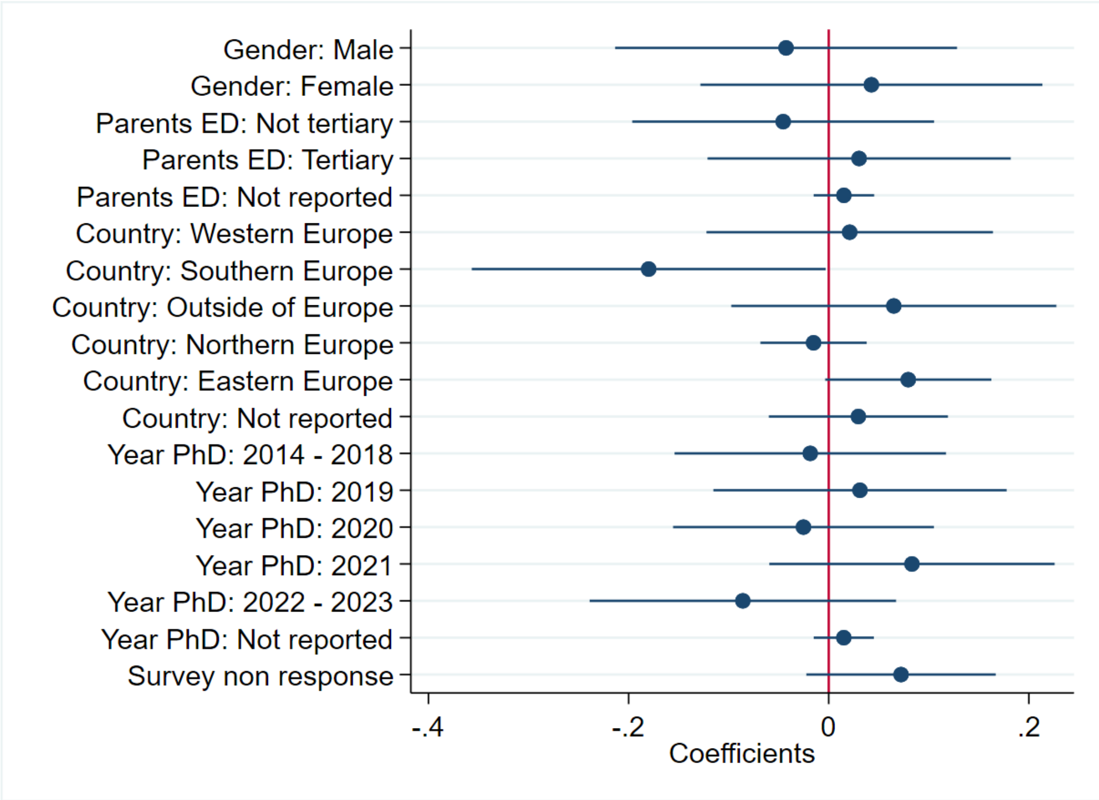
Notes: Table reports descriptive statistics at the reviewer level broken down by reviewer selectivity. Top 75 reviewers those from universities in the top 75 subject-specific QS ranking. It also includes mean differences and the corresponding p-values. The table highlights reviewers' demographic characteristics such as gender, parental education, and country of origin. We also gathered information on reviewers' preferences, rated on a scale from 1 to 10, regarding the following aspects of the conference submission process: peer review, anonymization of submissions, the inclusivity of PhD students and Postdocs, and the submission of extended abstracts. Additionally, reviewers rated the importance of several signals for themselves and the economics profession on a scale from 1 (Not important at all) to 5 (Extremely important). Finally, the table includes information on the number of conferences reviewers wish to attend and their willingness to network (rated from 1 (Not at all) to 5 (Very much)) through meetings or dinners with other economists or visiting scholars.

Table A.4: Selective institutions and position across rankings

Institution	QS	US News	Times HE	ShanghaiRanking	SCImago	RePEc
Harvard University	1	1	4	2	1	1
Stanford University	3	3	2	4	6	7
London School of Economics	7	7	10	8	2	19
University of Oxford	9	10	3	13	3	9
Columbia University	10	8	11	7	12	n/a
University of Cambridge	12	15	6	15	16	48
University of Pennsylvania	13	6	13	14	15	14
Bocconi University	16	20	n/a	32	59	96
University College London	17	41	n/a	17	10	18
University of Warwick	22	30	24	29	28	32
Boston University	22	23	54	27	67	16
Universitat Pompeu Fabra	26	124	85	51-75	206	n/a
Imperial College London	31	37	n/a	51-75	111	n/a
ETH Zürich	34	61	16	51-75	39	36
LMU Munich	38	126	47	51-75	94	76
Universität Mannheim	43	177	41	101-150	282	88
Erasmus University Rotterdam	45	10	22	34	31	59
Paris School of Economics	54	184	n/a	101-150	n/a	6
Barcelona School of Economics	62	n/a	n/a	76-100	n/a	12
Universitat Autònoma de Barcelona	62	143	101-125	101-150	163	222
University of Amsterdam	67	56	83	51-75	51	63
Université Paris 1 Panthéon Sorbonne	73	n/a	251-300	n/a	167	n/a
Ecole Polytechnique Paris	76	n/a	126-150	201-300	832	n/a
KU Leuven	77	46	62	51-75	54	51
Humboldt Uni. of Berlin	78	214	n/a	151-200	257	160
Queen Mary University of London	84	171	201-250	76-100	125	97
University of Bologna	85	107	151-175	101-150	70	40
Sciences Po	89	n/a	n/a	101-150	437	44
UCLouvain	91	197	176-200	151-200	245	n/a
Vrije universiteit Amsterdam	100	55	91	51-75	84	52
CEMFI	n/a	n/a	n/a	n/a	n/a	63
Bank of Italy	n/a	n/a	n/a	n/a	n/a	n/a

A.1.1 Balance checks

Figure A.1: Balance checks



Note: The figure reports 95% confidence intervals for balance checks. Results come from a series of regressions ran on a sample of applicants with the inclusion of a treatment indicator and pair FEs. Standard errors robust to heteroskedasticity.

A.2 Design

A.2.1 Example of Non-Visible Affiliation Extended Abstract

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis.

Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus.

Nulla facilisi. Etiam imperdiet imperdiet orci. Nunc nec neque. Phasellus leo dolor, tempus non, auctor et, hendrerit quis, nisi. Curabitur ligula sapien, tincidunt non, euismod vitae, posuere imperdiet, leo. Maecenas malesuada. Praesent congue erat at massa. Sed cursus turpis vitae tortor. Donec posuere vulputate arcu. Phasellus accumsan cursus velit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed aliquam, nisi quis porttitor congue, elit erat euismod orci, ac placerat dolor lectus quis orci.

A.2.2 Example of Visible Affiliation Extended Abstract

Author's Affiliation

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis.

Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus.

Nulla facilisi. Etiam imperdiet imperdiet orci. Nunc nec neque. Phasellus leo dolor, tempus non, auctor et, hendrerit quis, nisi. Curabitur ligula sapien, tincidunt non, euismod vitae, posuere imperdiet, leo. Maecenas malesuada. Praesent congue erat at massa. Sed cursus turpis vitae tortor. Donec posuere vulputate arcu. Phasellus accumsan cursus velit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed aliquam, nisi quis porttitor congue, elit erat euismod orci, ac placerat dolor lectus quis orci.

A.2.3 Grading Criteria

CATEGORY	EXPLANATION	Scale of answers
FILE NAME	Please input the file name for the submission you are evaluating	"file_name" from "file_name.pdf"
RESEARCH QUESTION	Please provide your mark based on: - The quality of the research question - How does the paper contribute to the relevant literature?	0 – 10
RESEARCH DESIGN	Please provide your mark based on: - Is the empirics/ theory well suited to provide an answer the research question of the paper? - To what extent are you convinced by the quality of the empirics/ theory presented in the paper?	0 – 10
WRITING/ PRESENTATION	Please provide your mark based on - How smoothly does the paper read?	0 – 10
OPEN-ENDED	Please provide a short comment explaining your evaluation	Open text
RECOMMENDATION	Would you recommend accepting this paper to the workshop?	(I) Definitely accept (II) Probably accept (III) Maybe accept (IV) I think this paper should not be accepted
RELATIVE QUALITY	Do you think this submission is of a higher/ lower/ about the same quality of an average paper from your PhD program?	(I) Higher (II) About the same (III) Lower
DISCUSSING RESEARCH	Would you be interested in discussing about research with the author of this submission?	(I) Yes (II) No

A.3 Choice of threshold for selectivity

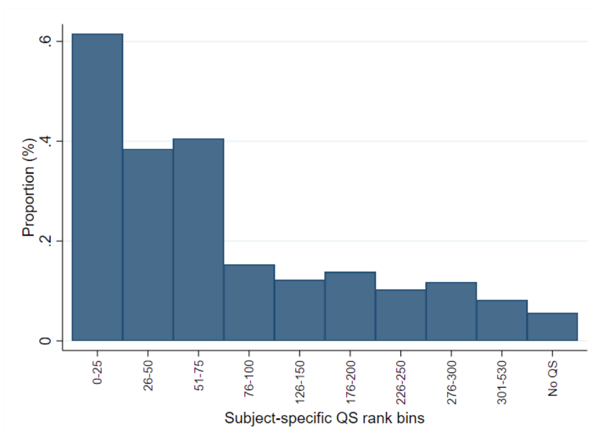
A.3.1 Results using reviewers' subjective classification of universities

Table A.5: Alternative specification using reviewers' subjective classification of universities

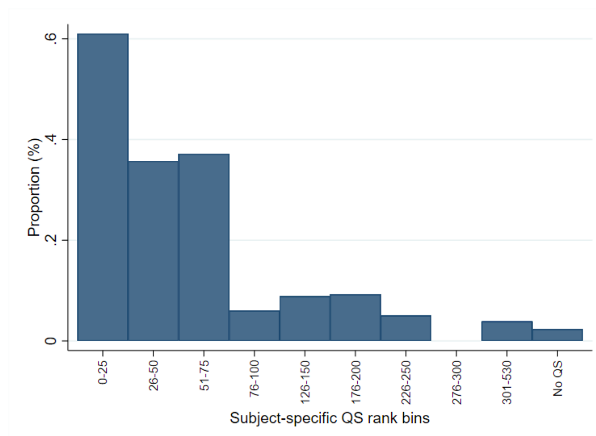
	(1) Overall score	(2) Position in ranking	(3) Suggestion for inclusion
VA	-0.47	1.43	-0.04
	(0.53)	(2.10)	(0.03)
VA × *Elite*	2.47***	15.48***	0.10
	(0.79)	(4.63)	(0.08)
N	899	905	905
Control mean	20.33	79.06	0.55

Note: The table reports β_1 and β_2 from Equation 1, substituting the *VA* × *Top75* interaction by an interaction of the treatment arm dummy and an indicator for whether the University of affiliation of the submitter was classified as "Elite" by the reviewer. The control mean refers to the average outcome in NVA grading. All specifications control for gender, year of start of the PhD, parental education, country of origin, and a stratification dummy for selective reviewers. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

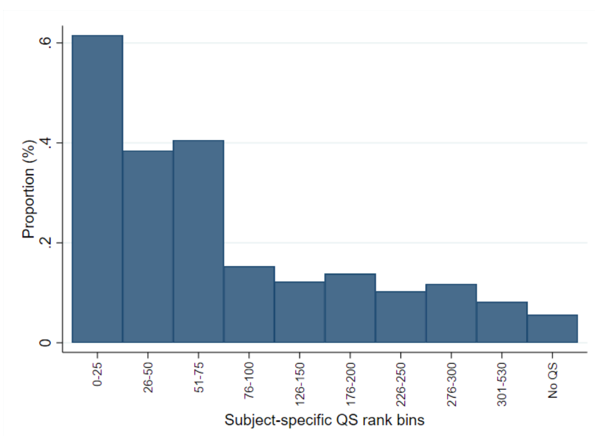
Figure A.2: Proportion of universities classified as "elite" by the reviewers in each QS rank bin.



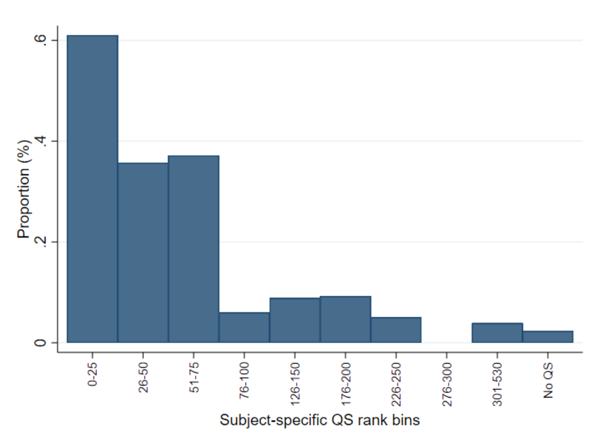
(a) 25 sized bins - Blind



(b) 25 sized bins - Non-blind



(c) 15 sized bins - Blind



(d) 15 sized bins - Non-blind

The figure reports the proportion of universities classified as "Elite" by blind and non-blind reviewers, plotted against the bin of the QS ranking where the university belongs. Panel (a) shows bins of size 25 positions in the ranking while Panel (b) refers to bins of size 15. The QS ranking refers to the subject-specific rank for Economics and Econometrics 2023.

A.4 Table of results by acceptance rate

Table A.6: Affiliation bias by acceptance rate

Panel A: Analysis for Acceptance Rate 10% to 40%						
	(1)	(2)	(3)	(4)	(5)	(6)
Acceptance rate:	10%	AEA 13%	20%	30%	40%	EEA 43%
VA	-0.02	-0.02	-0.04	-0.02	-0.02	-0.00
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
VA $\times$ *Top 75*	0.13***	0.13***	0.22***	0.11**	0.20***	0.18***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.06)	(0.06)
N	985	985	985	985	985	985
Control mean	0.11	0.11	0.23	0.28	0.37	0.43

Panel B: Analysis for Acceptance Rate 50% to 90%					
	(7)	(8)	(9)	(10)	(11)
Acceptance rate:	50%	60%	70%	80%	90%
VA	-0.00	-0.01	-0.01	-0.04***	-0.01
	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)
VA $\times$ *Top 75*	0.10**	0.11***	0.16***	0.16***	0.09***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)
N	985	985	985	985	985
Control mean	0.49	0.61	0.69	0.81	0.91

Note: Dependent variable refers to a dummy taking value 1 for submissions accepted according to different acceptance rates. The acceptance rate for the AEA ASSA meeting is 13% while the same figure for the EEA meeting is 43%. Estimates come from separate models, one for each acceptance threshold. The control mean refers to the average outcome in NVA grading. All specifications control for gender, separate dummies for PhD starting year, parental education, country of origin, a stratification dummy for selective reviewers and paper-by-block FEs. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution level.

A.5 Large Language Model Simulations

A.5.1 RCT Prompt

The following task is part of a research study. The purpose of the study is to learn how well you can predict the results of an experiment testing for affiliation bias in the peer review process of an early career workshop in economics. We define affiliation bias as any differential response to visible affiliation grading for TOP institutions. Your response will be used for academic research. In this experiment, applicants to an early career workshop in economics were required to review other submissions to the workshop. Submissions consisted of extended abstracts. The experiment has two treatment arms. In both arms, we removed from all submissions any identifying information linking authors to papers. Treatment A showed authors' affiliation only. Treatment B did not show affiliation. The two treatment arms differed only in terms of affiliation visibility. Reviewers were randomized across the two treatment arms and were not aware they are taking part in an experiment. We created couples of reviewers assigned to alternative treatment arms. Each couple received a set of 8 submissions. In each couple, one reviewer read 8 papers with visible affiliation and another reviewer read the same 8 papers without visible affiliation. We estimate treatment effects by comparing the same paper read by a reviewer who saw affiliation and a reviewer who did not see affiliation. This comparison removes concerns over reviewers grading on a curve. Each paper got assigned to multiple reviewers in both treatment arms. Reviewers were instructed to provide three grades from 0 to 10. Grade 1: How does the paper contribute to the relevant literature. Grade 2: Is the empirics/ theory well suited to provide an answer to the research question of the paper and to what extent are you convinced by the quality of the empirics/ theory presented in the paper. Grade 3: How smoothly does the paper read. The three grades were then summed into an overall score ranging from 0 to 30. Reviewers had to express a "Suggestion for inclusion" by selecting one of the following options: "Definitely accept", "Probably accept", "Maybe accept", and "I think this paper should not be accepted". The "Suggestion for inclusion" is a dummy equal to 1 if the reviewer chose "Definitely accept" or "Probably accept" and 0 for "Maybe accept" and "I think this paper should not be accepted". All reviewers were PhD candidates or Post-Docs. Here are the descriptives for reviewers in our sample: 37% of the reviewers were affiliated with a university in the top 100 QS subject specific economics ranking, 72% were male and 28% female, 36% were not european, 30% south european, 19% western european, 6% eastern european, and 6% unknown. 19% started their phd that same year or the one before, 22% started it 2 years before, 19% 3 years, 22% 4 years, and 18% more than 4, while 1% unknown. 62% of our reviewers came from tertiary educated families, 37% came from non tertiary educated households and 1% is unknown. β_1 is the effect of displaying affiliation for non-TOP institutions. It may take positive values if papers which are not from TOP institutions receive a premium or negative if they receive a penalty. β_2 is the coefficient of interest measuring any differential response to visible affiliation grading for TOP institutions. It may take on positive values if TOP universities receive a differential premium or a negative one if they receive a penalty. For Overall Score, these coefficients represent how much the grade changed, for Suggestion for inclusion they represent the change of probability of being suggested for inclusion. 29% of the reviewed papers come from a TOP institutions. We define "TOP" institutions as those in the top 75 positions of the QS subject-specific ranking for Economics and Econometrics. I want you to predict the results of this experiment with your best numeric predictions rounded to 2 decimals. Reply only by providing the coefficients in this order: β_1 Overall Score, β_1 Suggestion for Inclusion, β_2 Overall Score, β_2 Suggestion for Inclusion. Write only the predicted coefficients and ***, **, *, representing 1%, 5%, 10% significance levels respectively. Separate each estimate by ‘,’. Do not include any other text or label in your reply.

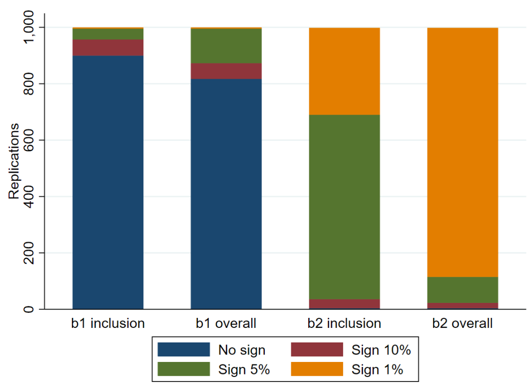
A.5.2 Out of sample Prompt

The following task is part of a research study. The purpose of the study is to learn how well you can predict the results of an experiment testing for affiliation bias in the peer review process of a workshop in economics. We define affiliation bias as any differential response to visible affiliation grading for TOP institutions. Your response will be used for academic research. In this experiment, applicants to a workshop in economics were required to review other submissions to the workshop. The experiment has two treatment arms. In both arms, we removed from all submissions any identifying information linking authors to papers. Treatment A showed authors' affiliation only. Treatment B did not show affiliation. The two treatment arms differed only in terms of affiliation visibility. Reviewers were randomized across the two treatment arms and were not aware they are taking part in an experiment. We created couples of reviewers assigned to alternative treatment arms. Each couple received a set of 8 submissions. In each couple, one reviewer read 8 papers with visible affiliation and another reviewer read the same 8 papers without visible affiliation. We estimate treatment effects by comparing the same paper read by a reviewer who saw affiliation and a reviewer who did not see affiliation. This comparison removes concerns over reviewers grading on a curve. Each paper got assigned to multiple reviewers in both treatment arms. Reviewers were instructed to provide three grades from 0 to 10. Grade 1: How does the paper contribute to the relevant literature. Grade 2: Is the empirics/ theory well suited to provide an answer to the research question of the paper and to what extent are you convinced by the quality of the empirics/ theory presented in the paper. Grade 3: How smoothly does the paper read. The three grades were then summed into an overall score ranging from 0 to 30. Reviewers had to express a “Suggestion for inclusion” by selecting one of the following options: “Definitely accept”, “Probably accept”, “Maybe accept”, and “I think this paper should not be accepted”. The “Suggestion for inclusion” is a dummy equal to 1 if the reviewer chose “Definitely accept” or “Probably accept” and 0 for “Maybe accept” and “I think this paper should not be accepted”. Both reviewers and applicants are senior economists from leading U.S. and European economics departments. Our reviewers are typically appointed to the editorial boards of prestigious economics journals. β_1 is the effect of displaying affiliation for non-TOP institutions. It may take positive values if papers which are not from TOP institutions receive a premium or negative if they receive a penalty. β_2 is the coefficient of interest measuring any differential response to visible affiliation grading for TOP institutions. It may take on positive values if TOP universities receive a differential premium or a negative one if they receive a penalty. For

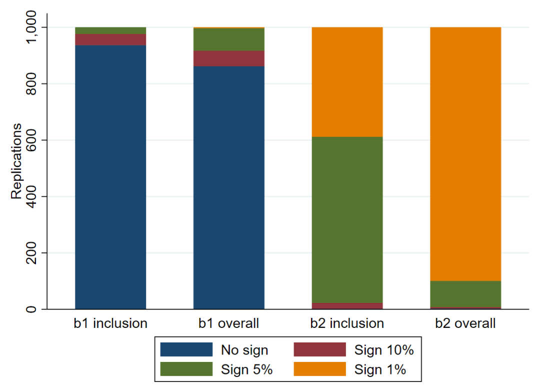
Overall Score, these coefficients represent how much the grade changed, for Suggestion for inclusion they represent the change of probability of being suggested for inclusion. We define "TOP" institutions as those in the top positions of the QS subject-specific ranking for Economics and Econometrics. I want you to predict the results of this experiment with your best numeric predictions rounded to 2 decimals. Reply only by providing the coefficients in this order: β_1 Overall Score, β_1 Suggestion for Inclusion, β_2 Overall Score, β_2 Suggestion for Inclusion. Write only the predicted coefficients and ***, **, *, representing 1%, 5%, 10% significance levels respectively. Separate each estimate by ','. Do not include any other text or label in your reply.

A.5.3 Predicted Significance

Figure A.3: Significance level by coefficient



(a) RCT sample prompt



(b) Out of sample prompt

Note: The graphs display the level of significance as predicted across simulations. Results are broken down by β_1 and β_2 coefficients for Overall score and Suggestion for inclusion.

A.6 Robustness Checks

Table A.7: Robustness check: Covariates and alternative fixed effects

	(1)	(2)	(3)
Panel A: Overall score			
VA	-0.48	-0.26	-0.14
	(0.55)	(0.53)	(0.45)
VA × *Top 75*	1.62***	1.22**	1.09*
	(0.55)	(0.59)	(0.57)
N	977	993	1103
Control mean	20.31	20.31	20.18
Couple p-value for Top 75			0.12
Covariates	Yes	No	Yes
Fixed Effects	Paper by Block	Paper by Block	Paper
Panel B: Position in ranking			
VA	2.87	2.93	3.63**
	(2.11)	(2.22)	(1.65)
VA × *Top 75*	18.59***	17.21***	16.91***
	(3.42)	(3.39)	(2.89)
N	985	1001	1107
Control mean	78.87	78.92	78.72
Couple p-value for Top 75			0.00
Covariates	Yes	No	Yes
Fixed Effects	Paper by Block	Paper by Block	Paper
Panel C: Suggestion for inclusion			
VA	-0.08*	-0.04	-0.06*
	(0.04)	(0.04)	(0.03)
VA × *Top 75*	0.20***	0.18***	0.16***
	(0.06)	(0.06)	(0.05)
N	985	1001	1107
Control mean	0.55	0.55	0.55
Couple p-value for Top 75			0.02
Covariates	Yes	No	Yes
Fixed Effects	Paper by Block	Paper by Block	Paper

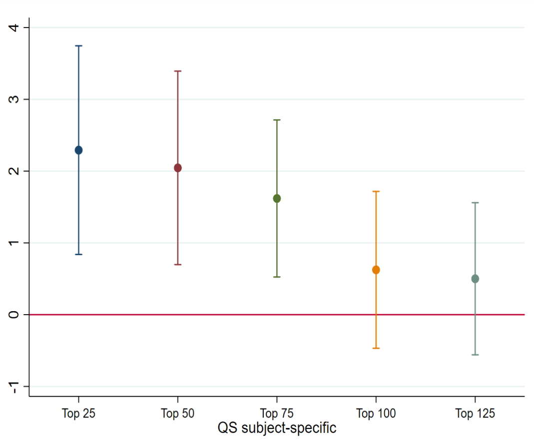
Note: Table replicates estimates for all outcomes in Table 1 across panels in column 1, removes covariates in column 2 and relaxes FEs from paper-by-block to paper level in column 3. The control mean refers to the average outcome in NVA grading. Covariates include gender, year of start of the PhD, parental education, and country of origin. All models control for a stratification dummy for selective reviewers. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution of the institution level. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

Table A.8: Multiple-hypothesis correction

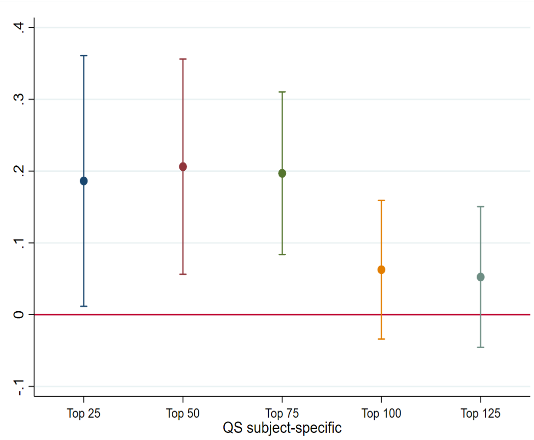
Outcome	Model	Resample	Holm	Bonferroni
Overall score - *VA*	0.39	0.57	0.57	1.00
Overall score - *VA* × *Top75*	0.00	0.00	0.01	0.02
Suggestion for inclusion - *VA*	0.05	0.19	0.56	1.00
Suggestion for inclusion - *VA* × *Top75*	0.00	0.01	0.03	0.03
Position in ranking - *VA*	0.18	0.37	0.73	1.00
Position in ranking - *VA* × *Top75*	0.00	0.00	0.00	0.00

Note: Table reports corrected p-values for six hypotheses of the estimates in Table 1 following Holm (1979); Bonferroni (1935). Resample p-values come from a bootstrap procedure with 1,000 replications with clustering at the institution level. Holm and Bonferroni p-values are computed from Resample p-values.

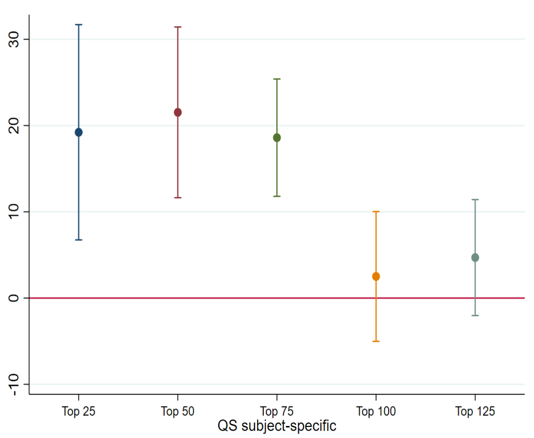
Figure A.4: Robustness of results to different thresholds for university selectivity



(a) Overall Score



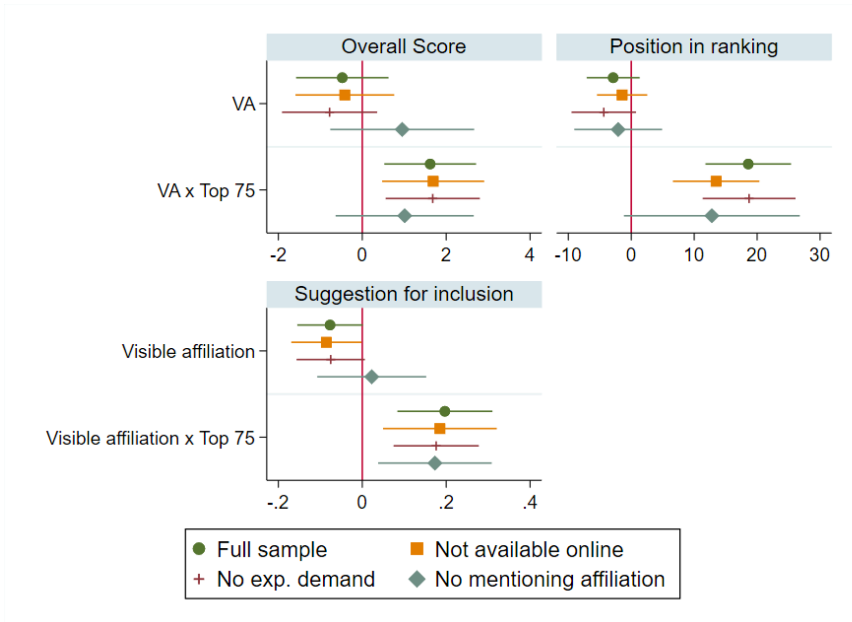
(b) Suggested for acceptance



(c) Position in the ranking

The figure reports 95% confidence intervals for separate estimates of β_2 from Equation 1, where we modify the variable of interest to reflect different definitions of selectivity.

Figure A.5: Original estimates, estimates after dropping observations at risk of violating treatment



The figure reports 95% confidence intervals for separate estimates of β_1 and β_2 from Equation 1. “Full sample” estimates refer to our main estimates from Table 1 while the label “Not available online” indicate the same parameters after dropping from the sample papers that could be found online, and hence attributable to an author. Estimates labelled “No mentioning affiliation” refer to results after dropping from the sample authors who mentioned affiliation in the open-ended section of the survey while estimates labelled “No exp. demand” come from dropping from the sample reviewers who mentioned an experiment.

A.6.1 QS methodology and sub-indicators

The QS World University Ranking by Subject is produced by Quacquarelly Symonds, a higher education analytics firm based in London. The 2023 version of the Economics and Econometrics ranking, the one used in this paper, considers 1,649 institutions and ranks the best 530 based on a weighted average of four indicators. The indicators are:

- Academic reputation (40%): a sample of over 130,000 academics are asked to list up to 10 domestic and 30 international universities that they consider to be excellent for research in the specific subject.
- Research citations per paper (20%): the indicator is constructed using citations data from Elsevier Scopus, after setting a discipline-specific minimum publication threshold and a weighting scheme to reflect the publication patterns.

- H-index (20%): in general, the H-index reflects the number of publications for which an author has been cited by other authors at least that same number of times. In this case, it is based on the author's most cited papers, and the number of citations they have received in other publications.

- Employer reputation (20%): a sample of more than 75,000 employers who hire graduate students are asked to identify up to 10 domestic and 30 international universities whose graduates they consider to be excellent for recruitment.

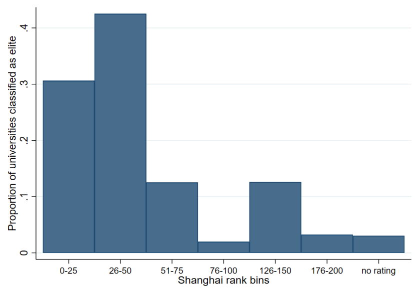
Table A.9: QS subject specific: Analysis by sub-indicator

	(1) Overall score	(2) Position in ranking	(3) Suggestion for inclusion
Panel A: Academic Reputation			
VA	-0.43	1.62	-0.07*
	(0.57)	(1.93)	(0.04)
VA × *Top 75*	1.17**	10.63***	0.13**
	(0.50)	(3.75)	(0.06)
Panel B: Employer Reputation			
VA	-0.58	3.38*	-0.07*
	(0.53)	(1.93)	(0.04)
VA × *Top 75*	2.63***	26.57***	0.21***
	(0.83)	(4.60)	(0.08)
Panel C: Citations			
VA	-0.32	0.40	-0.06
	(0.57)	(2.32)	(0.04)
VA × *Top 75*	0.99*	8.17**	0.14***
	(0.53)	(3.25)	(0.05)
Panel D: H index			
VA	-0.40	1.71	-0.07*
	(0.57)	(2.10)	(0.04)
VA × *Top 75*	1.17**	12.05***	0.15***
	(0.53)	(3.08)	(0.05)
N	977	985	985
Control mean	20.31	78.87	0.55

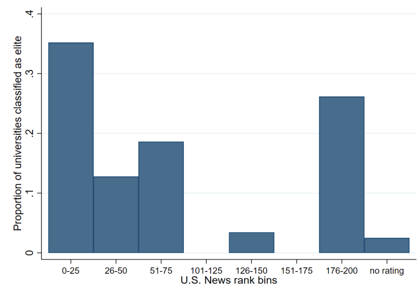
Note: Table reports estimates of β_1 and β_2 from Equation 1 using the different sub-indicators that build up the QS subject specific instead of the aggregate QS ranking position. The control mean refers to the average outcome in NVA grading. Across all panels and regressions we control for gender, year of start of the PhD, parental education, country of origin, and a stratification dummy for selective reviewers. Selective reviewers are those from universities in the top 100 subject-specific QS ranking. Standard errors are clustered at the institution. *, **, *** indicate statistical significance at 10%, 5%, and 1% level, respectively.

A.6.2 Robustness of results to different university rankings

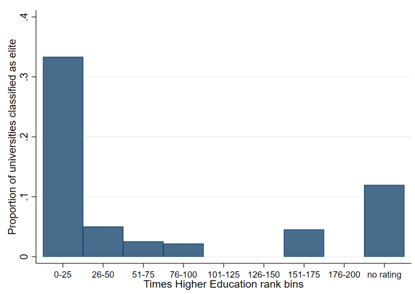
Figure A.6: Threshold choice



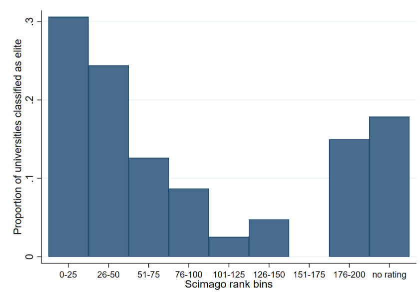
(a) ShanghaiRanking



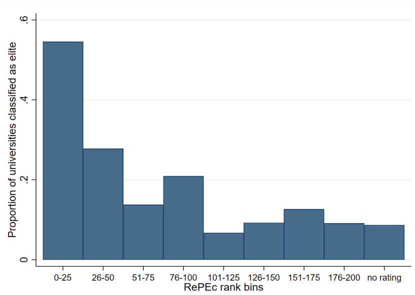
(b) U.S. News



(c) Times Higher Education



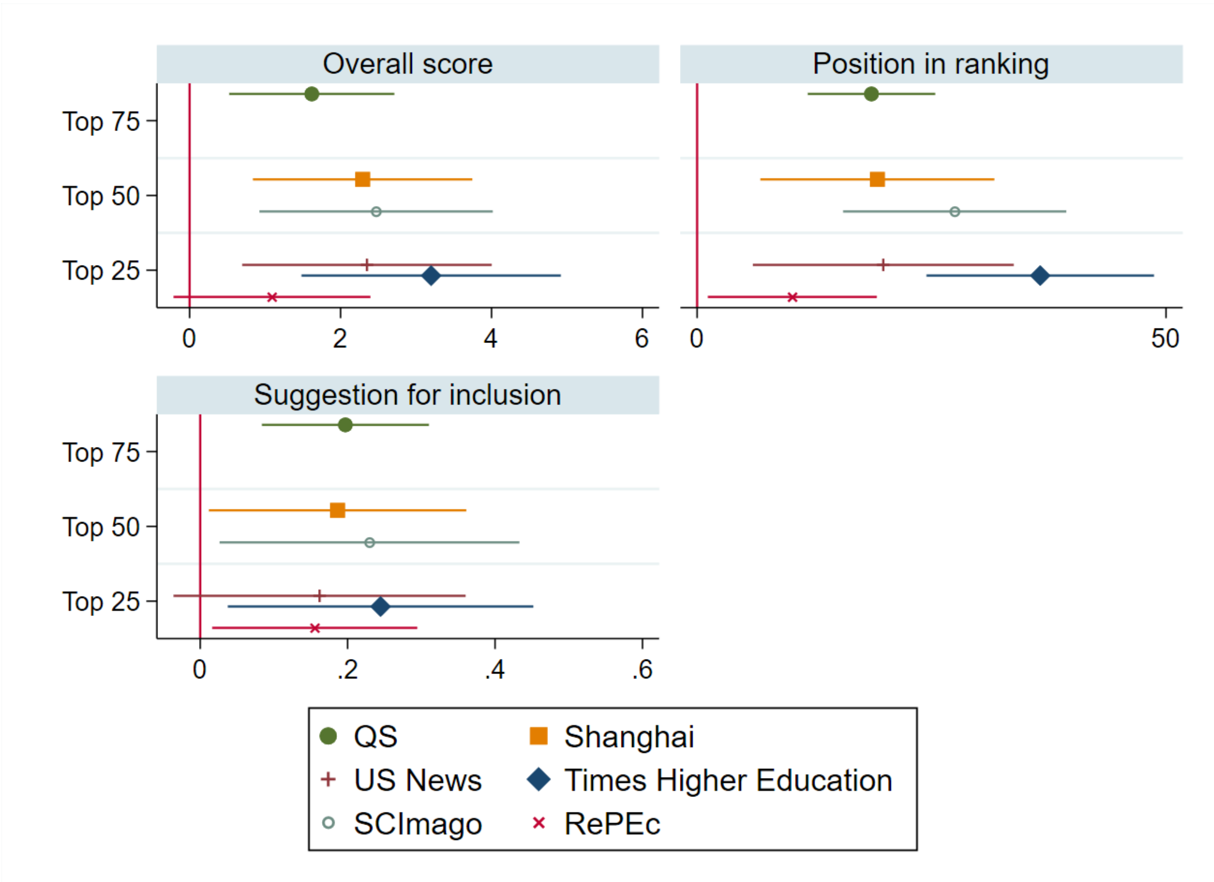
(d) SCImago



(e) RePEc

The figure reports the proportion of universities classified as “Elite” by the reviewers, plotted against the bin of the ranking where the university belongs.

Figure A.7: Main results across different university rankings



The figure reports 95% confidence intervals for separate estimates of β_2 from Equation 1, where we modify the variable of interest according to different institutional rankings. As shown in Figure A.6, we choose the specific threshold for each ranking by plotting the proportion of universities classified as "elite" by the reviewers against the bin of the ranking where the university belongs.

B Appendix B

B.1 Survey's Transcript

Notes for the reader: below is the transcribed text of the survey. Statements in [ ] are comments meant for the reader which were not included in the actual survey.

B.1.1 Welcome Page

Dear Applicant,

Welcome to the Grading Survey for the Ph.D. Workshop in Networks and Political Economy.

As previously arranged, you are requested to review 8 extended abstracts submitted to this workshop. We kindly point your attention to the instructions provided in the "Grading_criteria" file. If you encounter any questions or difficulties during the grading process, please do not hesitate to reach out to the organizing committee at **phd-workshop@univ-paris1.fr**.

We would like to emphasize that completing the **entire survey** is crucial. Failing to do so may lead to the **exclusion of your submission**, as it is an integral part of the grading process.

Completing the survey in its entirety should take about 10 minutes, as such we strongly recommend filling it out all at once. On the first page of the survey, you will find various questions, which can be filled out as explained in the "Grading_criteria" file.

You will find a set of 8 blocks, one for each submission you are requested to review. Each block requires you to input the file name referring to the abstract you are reviewing. You are free to input reviews in the order you choose.

IMPORTANT: every time you move to the next page you will not be able to come back, and your responses will be considered as definitive.

Your time and cooperation are greatly appreciated.

Sincerely,

The Organizing Committee

B.1.2 Page 1

1. **Paper 1**

   Please input the file name for the submission you are evaluating and the grades you awarded to the research question, research design, and writing/presentation quality. Provide your marks on a scale from 0 to 10.

2. **Open-ended Paper 1**

   Please provide a short comment explaining your evaluation

3. **Recommendation Paper 1**

   Would you recommend accepting this paper to the workshop?

   (a) Definitely accept

   (b) Probably accept

   (c) Maybe accept

   (d) I think this paper should not be accepted

4. **Relative Quality Paper 1**

   Do you think this submission is of a higher / lower / about the same quality of an average paper from your PhD program?

(a) Higher
(b) About the same
(c) Lower

5. **Discussing Research Paper 1**

   Would you be interested in discussing about research with the author of this submission?

   (a) Yes
   (b) No

[*These questions were repeated for each of the 8 papers under revision*]

B.1.3 Page 2

1. **How much did you like the following:**

	1	2	3	4	5	6	7	8	9	10
Peer review system										
Submission of extended abstracts instead of full papers										
Censoring of identifying information										
Having an early career conference open to both PhDs and Post Docs										

2. **What purpose do you think the peer review system serves?**

   Please write your answer here:

3. **What purpose do you think removing identifying information serves?**

   Please write your answer here:

4. **Is there anything else you would like to tell us?**

   Please write your answer here:

B.1.4 Page 3

1. **When reading a paper, in addition to the technical content, how important are the following elements to formulate your assessment?**

   (1=Not important at all; 5=Extremely important)

	1	2	3	4	5
Title					
Number of authors					
Gender of the author(s)					
Nationality of the author(s)					
Seniority of the author(s)					
Institutional affiliation of the author(s)					
Journal					
Acknowledgements					

B.1.5 Page 4

1. **When reading a paper, in addition to the technical content, how important do you think the economics profession considers the following elements?**

   (1=Not important at all; 5=Extremely important)

	1	2	3	4	5
Title					
Number of authors					
Gender of the author(s)					
Nationality of the author(s)					
Seniority of the author(s)					
Institutional affiliation of the author(s)					
Journal					
Acknowledgements					

B.1.6 Page 5

1. **How important do you think the following elements are for the career of a PhD candidate?**

   (1=Not important at all; 5=Extremely important)

	1	2	3	4	5
Attending workshops / conferences					
Having a good network					
Having a supervisor with a good network					

2. **How many workshops / conferences would you like to attend in a year?**

   (a) 0

   (b) 1

   (c) 2

(d) 3

(e) 4

(f) 5

(g) >5

3. **If given the opportunity, would you like to:**

   (1=Not at all, 5=Very much)

	1	2	3	4	5
Meet economists you do not know					
Have dinner with a visiting scholar who just arrived at your institution					

B.1.7 Page 6

1. **How would you assess the quality of the following research institutions?**

	Elite	Above average	Average	Below average	I don't know this institution
University A					
University B					
University C					
...					
University N					

B.1.8 Page 7

[*The questions of this section were not compulsory*]

This is the last section of the survey, and it will take very little time to fill out. Once you do, your application for the PhD Workshop in Networks and Political Economy @ Paris 1 Panthéon-Sorbonne will finally be official.

In this section of the survey, we are collecting a few pieces of information about you.

The information you provide will be used for a research project aimed at understanding how papers submitted for conferences are evaluated.

Please note that by submitting this form, you consent for your data to be used for research purposes only. Any individual data will be treated confidentially and anonymized.

This project has received ethical approval. For any information about the project, please contact: phd-workshop@univ-paris1.fr

Again, we thank you for your time and cooperation.

1. **Gender**
    (a) Female
    (b) Male
2. **What country are you from?**
    Please write your answer here:
3. **Please indicate your year of birth**
    Please write your answer here:
4. **What year did you begin your PhD?**
    Please write your answer here:
5. **What is the highest level of education completed by your parents?**

	Mother	Father
Less than high school		
High school		
Vocational training (post secondary)		
University degree (no PhD)		
PhD		
Prefer not to say		

C Online Appendix

C.1 Explanation of the different University Rankings

C.1.1 QS

The QS World University Ranking by Subject is produced by Quacquarelly Symonds, a higher education analytics firm based in London. The 2023 version of the Economics and Econometrics ranking, the one used in this paper, considers 1,649 institutions and ranks the best 530 based on a weighted average of four indicators. The indicators are:

- Academic reputation (40%): a sample of over 130,000 academics are asked to list up to 10 domestic and 30 international universities that they consider to be excellent for research in the specific subject.

- Research citations per paper (20%): the indicator is constructed using citations data from Elsevier Scopus, after setting a discipline-specific minimum publication threshold and a weighting scheme to reflect the publication patterns.

- H-index (20%): in general, the H-index reflects the number of publications for which an author has been cited by other authors at least that same number of times. In this case, it is based on the author's most cited papers, and the number of citations they have received in other publications.

- Employer reputation (20%): a sample of more than 75,000 employers who hire graduate students are asked to identify up to 10 domestic and 30 international universities whose graduates they consider to be excellent for recruitment.

There is no information available on the precise calculation of these indicators. As such, it is hard to assess its objectivity beyond the fact that the aggregated weight given to subjective components is 60%.

C.1.2 RePEc

Research Papers in Economics is a crowd-sourced initiative founded with the aim of facilitating the dissemination of research in economics. They use their publications database to produce rankings of journals, authors and institutions. The Top 25% Economics Departments ranking we use is based on the sum of the score of the authors affiliated with that department. Although this provides an advantage to big institutions, it is preferred to an average by the producers of the ranking because it provides an incentive to register all authors regardless of their rank. The authors' score is based on 31 different criteria, all based on publication metrics derived from the self-reported data contained in the RePEc repository, measuring things like number of works, citation counts, journal page counts and popularity on RePEc services. Different citation metrics are used, including raw count of citations, age-discounted, and four alternative measures weighted by Journal impact factor. For each author, the best and the worst ranking are dropped, and the 29 remaining are aggregated through a harmonic mean.

The main disadvantage in using this ranking comes from the fact that the ranking only considers Journals, authors and institutions registered in RePEc, which brings sample size down.

Although it is based on a transparent aggregation of 'objective' publication criteria, several papers show that the publication process and citations are also affected by multiple biases (Blank, 1991; Card et al., 2020; Carrell et al., 2024; Ersoy and Pate, 2023), so we expect to still find evidence of differential treatment using this ranking.

C.1.3 SCImago Institution Ranking

SCImago is a research group formed by members of the Spanish National Research Council (CSIC) and four Spanish universities - University of Granada, University of Extremadura, University Carlos III and University of Alcala de Henares. They produce analyses of the structure of the production of science, particularly the SCImago Institution Ranking, and the Atlas of Science. The SCImago Institutions Ranking evaluates academic and research-based institutions based on research, innovation and societal input criteria. In particular, these criteria are defined as follows:

- Research performance (50%): is a weighted average of different measures of citations, publications, publications in top 25% journals, publications in the top 10% most cited papers, percentage of output published as Open Access, number of authors, and research produced in collaboration with other institutuions.

- Innovation output (30%): it is an average of the amount of patents produced by the institution, and the patents that cite papers and patents produced by the institution.

- Societal impact (20%): is an aggregated measure of research output visibility in social media, Mendeley and policy documents, together with the number of female authors, the number of documents related to the Sustainable Development Goals and two metrics on the performance of the institution website.

The particularity of this ranking is that it includes criteria related to the impact of research in society and in innovation. While we do not expect the criterion based on patents to be particularly relevant in Economics, there are reasons to think Societal impact, in particular visibility in social media and internet presence, is of relevance in the discipline. We consider this criterion as of an intermediate subjectivity between the reputation-based criteria of the QS ranking, and the approach purely based on publications data used by RePEc.

https://www.scimagoir.com/methodology.php

C.1.4 Times Higher Education

Times Higher Education (THE) is a British magazine that reports news and issues on the topic of higher education. They produce university rankings since 2004, formerly in partnership with Quacquarelly Symonds, and from 2009 using data from Thomsom Reuters and a new methodology.

In order to be included in the ranking, universities must have produced more than one thousand relevant publications in the last five years preceding the year of publication of the ranking, and have a minimum of 5% or 50 academic staff working in the discipline. Moreover, universities that don't offer undergraduate education, or whose research output is concentrated (80% or more) in one subject area are excluded.

The ranking is based on the following measures:

- Teaching (30.4%): Teaching Reputation (estimated through a survey), staff-to-student ratio, doctorates-to-bachelor's ratio, doctorates awarded to academic staff ratio, and institutional income.

- Research environment (31.6%): Academic Reputation (also calculated through a survey among peer institutions), research income and number of publications normalized by institution size and subject.

- Research Quality (25%): citations, the 75th percentile of field-weighted citations, publications in the top 10 %, and citations weighted by the importance of the citing papers.

- International Outlook (9%): proportion of international students, proportion of international staff, and proportion of a university's total relevant publications that have at least one international co-author.

- Industry (4%): research income made by the university from industry, and number of patents that cite research conducted at the university.

C.1.5 Shanghai Ranking

The Academic Ranking of World Universities (ARWU), more widely known as the 'Shanghai Ranking' was first compiled and published by Shanghai Jiao Tong University, China in 2003, and since 2009 is published by the private firm ShanghaiRanking Consultancy. It is based on five indicators, namely:

- The number of journal publications in the first quartile of the Journal Impact Factor

- The ratio of citations of papers relative to the average citations in the same category and year

- The ratio of papers whose authors are located in at least two different countries relative to the total number of publications produced in the institution in the subject
- The number of papers published in top journals or top conferences (as defined by respondents to their ShanghaiRanking Academic Excellence Survey, which can be found in `https://www.shanghairanking.com/activities/aes`)
- The number of researchers that were awarded a significant award while employed with the institution (or whose last employer was the institution in case they were already retired) since 1991.

For the 2023 Economics ranking, the corresponding weights were 150, 50, 10, 100 and 100 respectively.

C.1.6 U.S. News Best Global Universities

U.S. News is a digital media company owned by Mortimer B. Zuckerman, co-founder, executive chairman and former CEO of Boston Properties, one of the largest real estate investment trusts in the U.S. It provides reporting, rankings, and advice in a variety of topics such as healthcare, personal finances, automobiles, real state, and education.

The 2022-2023 Best Global Universities for Economics and Business Ranking is based on seven publication and bibliometric criteria, provided by Clarivate and the Web of Science, together with two reputation and two international connections components, of which the source is not specified. The bibliometric indicators are based on the period 2016-2020, and the cutoff of citations is the 29th of May 2022.

In order to be included in the Economics and Business ranking, a University must have published a minimum of 250 papers in the field, bringing it to a total of 344 universities ranked in the Global version of the ranking.

The 11 indicators, with their respective weights in the final score, are:

- global research reputation (12.5%)
- regional research reputation (12.5%)
- publications (17%)
- normalized citation impact (7.5%)
- total citations (12.5%)
- number of publications that are amongst the 10% most cited (12.5%)
- percentage of total publications that are among the 10% most cited (5%)
- number of highly cited papers that are among the top 1% most cited in the field (5%)

- percentage of total publications that are among the top 1% most highly cited papers (5%)
- international collaboration relative to country (5%)
- unadjusted international collaboration (5%)

The variables are standardized to allow for intuitive interpretation and comparability of the different scales, and re-scaled to fall in a 0-100 range.

C.2 Comparison of the different University Rankings by completeness and popularity

Table C.1: Comparison of rankings used

Ranking	No. of unis in ranking	Threshold chosen	No. unis over threshold	General public popularity	Academics popularity
QS	68	75	19	2852	3080
ShanghaiRanking	65	50	10	2900	6190
US News	40	25	8	306	1750
Times Higher Education	67	25	8	1238	1630
SCImago	84	50	10	109	859
RePEc	49	25	8	38	203

C.3 Universities included in the "Top" category by ranking used

Table C.2: Comparison of "top" universities, by ranking

University	QS	RePEc	SCImago	Shanghai	THE	U.S. News	Total
Barcelona School of Economics	x	x					2
Bocconi University	x			x		x	3
Boston University	x	x		x		x	4
Columbia University	x		x	x	x	x	5
ETH Zürich	x		x		x		3
Harvard University	x	x	x	x	x	x	6
Imperial College London	x						1
LMU Munich	x						1
LSE	x	x	x	x	x	x	6
Paris School of Economics	x	x					2
Universitat Autonoma de Barcelona	x						1
Universitat Pompeu Fabra	x						1
University College London	x	x	x	x			4
University of Amsterdam	x						1
University of Birmingham			x				1
University of Cambridge	x		x	x	x	x	5
University of Oxford	x	x	x	x	x	x	6
University of Pennsylvania	x	x	x	x	x	x	6
University of Warwick	x	x	x	x			4
Universitat Mannheim	x						1