Research Article

Chang Chen[#], Jiayao Zhang[#], Ting Ye, Dan Roth, and Bo Zhang*

Causal inference with textual data: A quasi-experimental design assessing the association between author metadata and acceptance among ICLR submissions from 2017 to 2022

https://doi.org/10.1515/jci-2023-0052
received August 09, 2023; accepted July 22, 2024

Abstract: Many recent studies have probed status bias in the peer-review process of academic journals and conferences. In this article, we investigated the association between author metadata and area chairs' final decisions (Accept/Reject) using our compiled database of 5,313 borderline submissions to the International Conference on Learning Representations from 2017 to 2022 under a matched observational study framework. We carefully defined elements in a cause-and-effect analysis, including the treatment and its timing, pre-treatment variables, potential outcomes (POs) and causal null hypothesis of interest, all in the context of study units being textual data and under Neyman and Rubin's PO framework. We found some weak evidence that author metadata was associated with articles' final decisions. We also found that, under an additional stability assumption, borderline articles from high-ranking institutions (top-30% or top-20%) were less favored by area chairs compared to their matched counterparts. The results were consistent in two different matched designs (odds ratio = 0.82 [95% confidence interval (CI): 0.67 to 1.00] in a first design and 0.83 [95% CI: 0.64 to 1.07] in a strengthened design) and most pronounced in the subgroup of articles with low ratings. We discussed how to interpret these results in the context of multiple interactions between a study unit and different agents (reviewers and area chairs) in the peer-review system.

Keywords: matched observational study, natural language processing, peer-review, quasi-experimental design, status bias

MSC 2020: 62A01, 62P25

# The first two authors contributed equally to the work.

*** Corresponding author: Bo Zhang**, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, WA 98109, United States, e-mail: bzhang3@fredhutch.org
Chang Chen: Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27514, United States, e-mail: waltchen@unc.edu
Jiayao Zhang: Department of Computer and Information Science and Department of Statistics and Data Science, University of Pennsylvania, PA 19104, United States, e-mail: jiayaozhang@ac.org
Ting Ye: Department of Biostatistics, University of Washington, WA 98195, United States, e-mail: tingye1@uw.edu
Dan Roth: Department of Computer and Information Science, University of Pennsylvania, PA 19104, United States, e-mail: danroth@seas.upenn.edu

1 Introduction

1.1 Peer-review process

Peer-review has been the cornerstone of scientific research. It is important that the peer-review process be fair and impartial, especially for early-career researchers. In recent years, the peer-review process has been under a lot of scrutiny. For instance, in 2014, the organizers of the Conference on Neural Information Processing Systems (NeurIPS) randomly duplicated 10% of the submissions and assigned them to two independent sets of reviewers. The study found that 25.9% of these submissions received inconsistent decisions. Cortes and Lawrence [1] tracked the fate of submissions rejected in the NeurIPS experiment and found that the peer-review process was good at identifying poor papers but fell short of pinpointing good ones. McGillivray and De Ranieri [2] analyzed 128,454 articles in Nature-branded journals and found that authors from less prestigious academic institutions are more likely to choose double-blind review as opposed to single-blind review. To further illuminate the difference between single-blind and double-blind review processes, Tomkins et al. [3] assigned each submission from 2017 Web Search and Data Mining conference to two single-blinded reviewers and two additional double-blinded reviewers, and they found single-blind reviewing conferred a significant advantage to papers with authors from high-prestige institutions. More recently, Sun et al. [4] studied 5,027 papers submitted to the International Conference on Learning Representations (ICLR) and found that scores given to the most prestigious authors significantly decreased after the conference switched its review model from single-blind review to double-blind review. Smirnova et al. [5] evaluated a policy that encouraged (but did not force) authors to anonymize their submissions and found that the policy increased positive peer-reviews by 2.4% and acceptance by 5.6% for low-prestige authors and slightly decreased positive peer-reviews and acceptance rate for high-prestige authors. Many of these studies identified associations between decision makers' perception of certain aspects of articles' author metadata (e.g., authors' prestige or identity) and final acceptance decisions of these articles, and suggested various forms of implicit bias in the peer-review processes, especially among those that adopt a single-blind model.

1.2 Hypothetical experiment

In a seminal paper, Bertrand and Mullainathan [6] measured racial discrimination in labor markets by sending resumes with randomly assigned names, one African American sounding and the other White sounding (e.g., Lakisha versus Emily), to potential employers. Bertrand and Mullainathan's [6] study was elegant for two reasons. First, it was a randomized experiment free of confounding bias, observed or unobserved, although to what extent the found effect could be attributed to the bias toward applicants' race and ethnicity versus toward other personal traits signaled by the names is unclear (see, e.g., related discussions in Bertrand and Mullainathan [6, Section IV] and Greiner and Rubin [7]). Second, the study illustrated a general strategy to measure the causal effect due to an immutable trait: instead of imagining manipulating the immutable trait itself, the study manipulated employers' *perception* of this immutable trait.

In a recent high-profile study published in the *Proceedings of the National Academy of Sciences*, Huber et al. [8] designed a field experiment in the similar spirit as Bertrand and Mullainathan [6]. Huber et al. [8] measured the extent of the *status bias*, defined as a differential treatment of the same paper by prominent versus less established authors in the peer-review process, by randomizing over 3,300 researchers to one of the three arms: one arm assigned an article with a prestigious author, one arm assigned an anonymized version of the same article, and the other arm assigned the same article but with a less established author. Huber et al. [8] found strong evidence that the prominence of authorship markedly increased the acceptance rate by as much as sixfold. Using a similar study design, Nielsen et al. [9] examined the extent to which country- and institution-related status bias in six disciplines, namely, astronomy, cardiology, materials science, political science, psychology, and public health, and found inconclusive evidence supporting a status bias.

Bertrand and Mullainathan [6] and Huber et al.'s [8] studies illuminate a randomized experiment that conference organizers and journal editorial offices could carry out, at least hypothetically, in order to understand various forms of bias. For instance, if the policy interest is to evaluate the effect of reviewers' perception of certain aspects of authors (e.g., authors' identity and institution), then a hypothetical experiment would forward to reviewers articles with randomly assigned aspects of interest. Although such an experiment is conceivable, it is difficult to implement due to practical constraints.

1.3 Quasi-experimental design using a corpus of ICLR papers

In the absence of a randomized controlled trial (RCT) a quasi-experimental design aims to fill in the gap by constructing two groups, one treatment group and the other comparison group, that are as similar as possible in pre-treatment variables from retrospective, observational data. Statistical matching is a popular quasi-experimental design device [10,11]. In this article, we describe an effort to conduct a matched observational study that investigates the effect of authorship metadata on articles' final decisions using state-of-the-art natural language processing (NLP) tools and quasi-experimental design devices.

Our database was constructed from a carefully curated corpus of articles from ICLR, a premium international machine learning (ML) conference. The database is the first of its kind to provide an empirical evidence base for investigating the peer-review process. In particular, the database is feature-rich, in the sense that it contains not only explicit/structural features of an article such as its keywords, number of figures, and author affiliations, but also more subtle and higher-level features such as topic and textual complexity as reflected by articles' text, and reviewers' sentiment as reflected by their publicly available comments. Building upon Neyman and Rubin's potential outcomes (PO) framework [12,13] and discussions of immutable traits regarding human subjects, for instance, those in Greiner and Rubin [7], we elaborate on the essential elements of causal inference that facilitate a cause-and-effect analysis between authorship and papers' final decisions; in particular, we will carefully define and state the treatment of interest including its timing, pre-treatment variables, causal identification assumptions, causal null hypothesis to be tested, and how to interpret the results, all in the context of study units being textual data.

The conference submission and peer-review process consists of multiple steps. For a typical ML conference such as ICLR, articles need to be submitted by authors before a pre-specified deadline. Valid submissions are then forwarded to a number of reviewers (typically three to four) for feedback and a numerical rating. This part of the peer-review process is double-blind so the reviewers and authors in principle do not know each other, although in practice, reviewers could identify authors from penmanship or because the authors might have uploaded their articles to the preprint platform arXiv.org. Authors are then given the chance to answer reviewers' comments and feedback and provide a written rebuttal. Reviewers are allowed to modify their previous ratings taking into account the rebuttal. Finally, an area chair (similar to an associate editor of an academic journal) reviews the article and its ratings and then makes a final decision. Submitted articles, author metadata, reviewers' written comments, authors' written rebuttals, reviewers' ratings and area chairs' final decisions are all openly available from the website openreview.net.

Our compiled database allows us to study many different aspects of the peer-review process. In this article, we will focus specifically on the last stage of the peer-review process and investigate the effect of authorship metadata on area chairs' final decisions. Our focus was motivated by several considerations. First, it is an empirical fact that articles receiving identical or near-identical ratings could receive different final decisions (see, e.g., the stacked bar graph in Panel a of Figure 1). It is not unreasonable for authors, especially those who are junior and less established, to wonder if they are fairly treated. Second, any endeavor to establish a cause-and-effect relationship using observational data is challenging because of unmeasured confounding. In our analysis, articles have many implicit features such as novelty and thoughtfulness, and these unmeasured confounders could, in principle, explain away any association between author metadata and final decisions. This problem is greatly attenuated when we focus on area chairs' final decisions and have reviewers' ratings

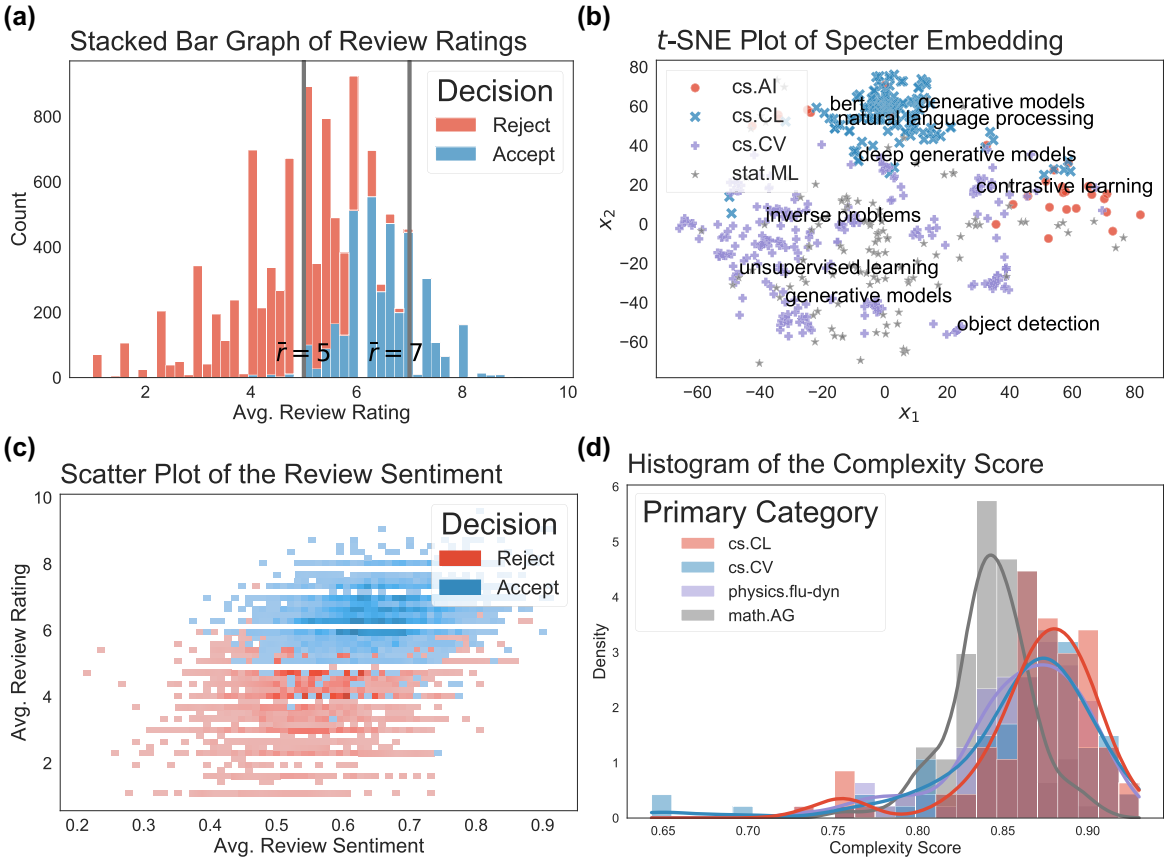

Figure 1: Illustration of ratings and high-level textual features: (a) stacked bar graphs of acceptance decisions. The decisions of borderline articles (average rating between 5 and 7) are not clearcut, (b) T-SNE plot of the Specter embedding together with arXiv primary categories and primary keywords among submissions that have been put onto arXiv, (c) sentiment and ratings of reviews, and (d) distribution of complexity scores of a random samples of arXiv articles across four primary categories.

as *pre-treatment* covariates: reviewers' ratings should be a good, albeit not perfect, proxy for the innate quality of an article.

The plan for the rest of the article is as follows. In Section 2, we briefly describe our compiled database. In Section 3, we lay out the PO framework under which we will conduct the matched observational study. Section 4 describes two concrete study designs. We also report outcome analysis results in Section 4. Section 5 discusses how to interpret our findings.

2 Data: ICLR papers from 2017–2022

2.1 Database construction and structural features

We used the ICLR database[1] collected and complied by Zhang et al. [14]. Motivated by the observation that area chairs' final decisions of submissions with an average rating between 5 and 7 are not deterministic, as shown

1 https://cogcomp.github.io/iclr_database/

in Panel a of Figure 1, we restricted ourselves to this subset of borderline submissions. We first briefly recall the data collection and cleaning process here for completeness.

The `OpenReview` API was used to crawl data of 10,289 submissions to the ICLR from 2017 to 2022. The crawled data include (i) submissions and their metadata, (ii) author metadata, (iii) review/rebuttal data and (iv) area-chair/program chair decision. Structural features, including the number of sections, figures, tables, and bibliographies, were extracted in a relatively straightforward manner. We also extracted and ranked self-reported keywords from each submission to form primary and secondary keywords. Author profiles include an optional self-reported gender; we used the first name dictionary developed by Tsai et al. [15] to provide a numerical score based on the first names of the authors, where 0 signifies female and 1 otherwise. Author profiles were then augmented via Google Scholar API to obtain author citation and h-index data. Author institution was matched using the domain name of the author email.[2] Although CSRanking data are available, it does not have full coverage of all authors' institutions. As such, we mainly used the institutional ranking derived from the cumulative number of accepted papers to the ICLR in the past. For example, the ranking in 2020 of institution A was determined by all papers accepted to ICLR 2017–2020 that had at least one author from it. The review data include rating, confidence, and textual reviews. In some years, for example, 2020 and 2022, there were additional assessments such as technical soundness or novelty. Since these additional assessments were not available for all years, we restricted our attention to ratings, confidence, and higher-level features derived from textual reviews to be discussed shortly. Finally, we dichotomized the paper decision by grouping various acceptance designations (spotlight, poster, short talk) into "Accept" and "Reject" or "invited to workshop track" as "Reject."

We also identified if a submission was posted on the preprint platform arXiv.org before the review deadline by (i) searching for five most relevant results based on the title and abstract corresponding to each article from arXiv.org, (ii) computing the Jaccard similarity and normalized Levenshtein similarity between authors, and (iii) calculating the cosine-similarity of the title-abstract embedding. Using the arXiv timestamp, we then identified which submissions were posted prior to the end of the review process. Among the subset of papers that had arXiv counterparts, we also obtained their arXiv-related metadata such as primary and secondary categories.

2.2 Derived higher-level features

Although the structural features described so far contain abundant information, we considered further leveraging language models to extract additional higher-level features directly from textual data. These higher-level features, such as topics, quality of writing and mathematical complexity, or rigor, may help quantify residual confounding not captured by structural aspects (e.g., those described in Section 2.1) of an article. Furthermore, in a matched observational study, it is desirable to have a characterization of the "similarity" among study units to serve as a criterion for matching. Therefore, we derived the following higher-level features and a similarity measure based on embeddings from language models to facilitate our study design.

2.2.1 SPECTER embedding

Our first tool is the SPECTER model [16], a bidirectional encoder representations from transformers (BERT)-based model [17] fine-tuned on a scholarly corpus that has a good track record on summarization tasks (generate abstracts from the main texts) on academic articles. This model takes the *abstract* of the submissions and outputs a 784-dimensional real-valued vector as its representation. Panel b of Figure 1 plots a two-dimensional t-distribution stochastic neighbor embedding (t-SNE) [18] embedding of this representation across

2 Only domain names are visible from the OpenReview API; other information is masked.

a subset of 535 submissions that have their arXiv information and primary keyword available. We see that computer vision and computational linguistics articles separate well while general artificial intelligence (AI) and ML articles blend in. In addition, we sample nine primary keywords to overlay on the *t*-SNE embedding. Note that semantically similar keywords (e.g., the phrases "BERT" and "natural language processing") generally have higher proximity under this embedding, which further demonstrate its effectiveness. We thus used this embedding to (1) perform a ten-component spectral clustering to assign each submission a "semantic cluster" of submissions, and (2) compute the cosine similarity between any two articles.

2.2.2 Sentiment modeling

A RoBERTa model [19] fine-tuned on Twitter sentiments [20] was used to assign a sentiment score to textual reviews, where 0 signifies negative and 1 positive. We plot the scatter plot of average sentiment and average rating of submissions with different color signifies different decisions in Panel c of Figure 1. We observe that the sentiment is highly correlated with the rating while behaving more volatile when the rating is borderline. This suggests that incorporating review sentiment may help complement numerical ratings in the downstream analysis, especially when numerical ratings are borderline and not discriminative.

2.2.3 Complexity score

We used an off-the-shelf fluency model[3] derived from the RoBERTa model to assess sentence-level fluency (1 signifies most fluent and easy-to-read, while 0 denotes gibberish-like sentences). We then took the average to represent the complexity of an article. This fluency score measures how well each article aligns with the English grammar and serves as a proxy for an article's heaviness in mathematical notation since in-line mathematical notation often disrupts an English sentence's fluency and results in a lower score. In Panel d of Figure 1, we perform a sanity-check by randomly sampling approximately 100 arXiv papers from four categories (computational linguistics, computer vision, fluid dynamics, and algebraic geometry) and compute their complexity score. Most of the scores are relatively high, as expected since academic articles are often relatively well written. We also observe a discrepancy in that algebraic geometric papers have their score distribution significantly skewed to the left, which also aligns with our intuition. We thus used this complexity score as a proxy to mathematical complexity and paper readability in our subsequent analysis.

3 Notation and framework

3.1 Matched cohort study

In the absence of a well-controlled experiment (e.g., the hypothetical RCT envisioned in Section 1.2), observational studies, *up to their important limitations*, provide an alternative to explore a cause-and-effect relationship [21]. In an observational study, study units receiving different levels of treatment may differ systematically in their observed covariates, and this induces the so-called *overt bias* [10, Section 3]. In our case study, articles with different author metadata, for instance, those with authors from high-ranking versus relatively lower-ranking academic institutions, could differ systematically in topics, keywords, number of figures, and equations, among others, and this would invalidate a naïve comparison.

Statistical matching is a commonly used strategy to adjust for confounding in empirical studies [10,11,22–25]. The goal of statistical matching is to embed non-randomized, observational data into an

3 https://huggingface.co/prithivida/parrot_paraphraser_on_T5

approximate randomized controlled experiment by *designing* a matched control (or comparison) group that resembles the treated group in observed pre-treatment covariates by matching on these covariates [22,23], a balancing score derived from these covariates, e.g., Rosenbaum and Rubin's [26] propensity score, or a combination of both [27].

We note that there are multiple ways to adjust for observed covariates and draw causal conclusions under the PO framework, statistical matching being one of them. Other commonly used methods include weighting, modeling the PO, and a combination of both. We found a matched observational study particularly suited for our case study for three reasons. First, it facilitates testing Fisher's sharp null hypothesis of no effect, which is an appropriate causal null hypothesis encoding an intriguing notion of *fairness*, as we will discuss in detail in Section 3.4. Second, a matched design naturally takes into account similarity of textual data (for instance, as measured by cosine similarity based on their embeddings) and is capable of balancing some high-dimensional covariates like keywords in our data analysis. A third strength is mostly stylistic: a matched comparison best resembles Bertrand and Mullainathan's [6] seminal field experiment and is perhaps the easiest-to-digest way to exhibit statistical analysis results to a non-technical audience.

In the rest of this section, we articulate essential elements in our analysis, including study units, treatment to be hypothetically manipulated, PO, timing of the treatment, pre-treatment variables, and causal null hypothesis of interest.

3.2 Study units; treatment and its timing; POs

As discussed in detail in Greiner and Rubin [7], there are two agents in our analysis of the effect of authorship metadata on area chairs' final decisions: an ICLR article peer-reviewed and having received reviewers' ratings, and a decision maker, i.e., an area chair or meta-reviewer (AC), who assigned the final acceptance status to the article. In our analysis, each study unit is a (peer-reviewed ICLR article, area chair) pair. There are a total of N = 10,289 study units in our compiled database, and 5,313 of them have three or four reviewers and an average rating between 5 and 7. We will write the ith study unit as SU_i = (article$_i$, AC$_i$).

We define the treatment of interest as an area chair's perception of a peer-reviewed article's authorship metadata. This definition is modeled after Bertrand and Mullainathan [6] and Greiner and Rubin [7] and implies the *timing* of the treatment: we imagine a hypothetical randomized experiment where peer-reviewed ICLR articles, whose constituent parts include text, tables, figures, reviewers' ratings and comments, are randomly assigned authorship metadata and presented to the area chair for a final decision. This timing component of the treatment is critical because it implies which variables are "pre-treatment variables" under Neyman–Rubin's causal inference framework, as we will discuss in detail in Section 3.3.

In principle, the most granular author metadata is a complete list of author names with their corresponding academic or research institutions. Let $\vec{A} = \vec{a}$ denote author metadata and $\mathcal{A}$ the set of all possible configurations of author metadata. There is one PO $Y_i(\vec{a})$ associated with unit i and each $\vec{a} \in \mathcal{A}$; in words, there is one final decision associated with each peer-reviewed article had the author metadata been $\vec{a}$. We will assume the consistency assumption so that the observed outcome $Y_i^{\text{obs}} = Y_i(\vec{a}^{\text{obs}})$. One may adopt a variant of the *stable unit treatment value assumption* or SUTVA [28] to reduce the number of POs. For instance, one may further assume that the PO $Y_i(\vec{a})$ depends on author metadata $\vec{a}$ only via authors' academic institutions. Let $f(\cdot)$ denote a mapping from author metadata to authors' academic institutions, then this "stability" assumption amounts to assuming $Y_i(\vec{a}) = Y_i(\vec{a}')$ when $f(\vec{a}) = f(\vec{a}')$. We do not *a priori* make such stability assumptions.

Example 1. (Field experiment in Bertrand and Mullainathan [6]) In Bertrand and Mullainathan [6] field experiment, each study unit consists of a resume i and a human resource person reading the resume i, i.e., SU_i = (resume$_i$, HR person$_i$). Treatment is a person's perception of the name on the resume. In this case, $\mathcal{A}$ would consist of all names and $Y_i(A = a)$ is the potential administrative decision had the resume i been

associated with name a. If we further make the stability assumption that $Y_i(a)$ depends on a only via its race and ethnicity connotation as in Bertrand and Mullainathan [6] and define $f(a) = 1$ if the name a is African-American sounding and 0 if it is White sounding, then the set of POs $Y_i(a)$, $a \in \mathcal{A}$ would reduce to $\{Y_i(1), Y_i(0)\}$.

3.3 Observed and unobserved pre-treatment variables

According to Rubin [29], covariates refer to "variables that take their values before the treatment assignment or, more generally, simply cannot be affected by the treatment." In the following, we briefly review a dichotomy of pre-treatment variables in the context of drawing causal conclusions from textual data [30].

In human populations research, pre-treatment variables or covariates are often divided into two broad categories: observed and unobserved (see, e.g., Rosenbaum [10,11]). A randomized controlled experiment like the one in Bertrand and Mullainathan [6] had the key advantage of balancing both observed and unobserved confounding, while drawing causal conclusions from observational data inevitably suffers from the concern of unmeasured confounding and researchers often control for a large number of observed covariates in order to alleviate this concern.

When study units are textual data, Zhang and Zhang [30] divided observed covariates into two types: *explicit observed covariates* $\mathbf{X}_{\text{obs}}^{\text{exp}}$ that could be derived from textual data at face value, e.g., number of equations, tables, and illustrations in the article, and *implicit observed covariates* $\mathbf{X}_{\text{obs}}^{\text{imp}}$ that capture higher-level aspects of textual data, e.g., the topic, flow, and novelty of the article. In our case study, we will consider the following explicit observed covariates: year of submission, reviewers' ratings, number of authors, sections, figures and references, and keywords. We will also consider each article's complexity, topic and reviewers' sentiment extracted using the state-of-the-art, NLP models as described in Section 2.

Unmeasured confounding is a major concern for any attempt to draw a cause-and-effect conclusion from observational data, regardless of the covariance adjustment method. Despite researchers' best intention and effort to control for all relevant pre-treatment variables via matching, there is always a concern about unmeasured confounding bias as we are working with observational data. In our analysis of ICLR papers, we identified two sources of unmeasured confounding. First, there could be residual confounding due to the insufficiency of language models (such as the SPECTER model) in summarizing or extracting implicit observed covariates $\mathbf{X}_{\text{obs}}^{\text{imp}}$ such as topics, flow, and sentiment. Second, in our analysis, we used numeric ratings from reviewers as a proxy of the quality and novelty of the article. Reviewers' ratings may not be sufficient in summarizing the quality of the articles. Unmeasured confounding may lead to a spurious causal conclusion, and researchers routinely examine the robustness of the putative causal conclusion using a sensitivity analysis [10,11,31].

3.4 Causal null hypothesis: A case for Fisher

A causal statement is necessarily a comparison among POs. In the context of a many-level treatment assignment, Fisher's sharp null hypothesis states the following:

$$H_{0,\,\text{sharp}}: \quad Y_i(\vec{a}) = Y_i(\vec{a}'), \quad \forall i = 1, \ldots, N, \quad \text{and} \quad \vec{a}, \vec{a}' \in \mathcal{A}. \tag{1}$$

Fisher's sharp null hypothesis prescribes a notion of fairness that, arguably, best suits our vision: area chairs' final decisions of the articles are *irrelevant* of author metadata; in other words, the decision $Y_i(\vec{a})$ could potentially depend on any substantive aspect of the article i, including its topic, quality of writing, and reviewers' ratings, but would remain the same had we changed author metadata from $\vec{a}$ to $\vec{a}'$.

In addition to Fisher's sharp null hypothesis, Neyman's weak null hypothesis, which states that the sample average treatment effect is zero, is another commonly tested causal null hypothesis. Unlike Fisher's sharp null,

Neyman's weak null hypothesis allows perception bias of varying magnitude for all article-AC pairs, as long as these biases would cancel out each other in one way or another. We found this a sub-optimal notion of fairness compared to that encoded by Fisher's sharp null hypothesis, and we will focus on testing Fisher's sharp null hypothesis in our data analysis.

Example 2. (continuing from p. 7) Bertrand and Mullainathan [6] found that White sounding names receive 50% more callbacks for interviews; under the stability assumption discussed in Section 3.2, their findings could be interpreted as a causal effect of perceiving White versus African-American sounding names. In the absence of the stability assumption, Bertrand and Mullainathan's result could still be interpreted as providing evidence against Fisher's sharp null hypothesis $H_{0,\text{sharp}}$ in its most generic form, although in what specific ways $H_{0,\text{sharp}}$ is violated needs further investigation.

Unlike Bertrand and Mullainathan's [6] field experiment that randomly assigned resume names, our cohort of ICLR articles are not randomly assigned authorship metadata. It is conceivable that articles with more "prestigious" authors, however one might want to define the concept of "prestige," could differ systematically in their reviewers' ratings, topics, etc., and this difference in baseline covariates could potentially introduce a spurious association between author metadata and area chairs' final decisions. To overcome this, we embed the observational data into a matched-pair design by constructing I matched pairs, each with two peer-reviewed articles, indexed by $j = 1, 2$, such that these two articles are as similar as possible in their covariates but with different author metadata. Let $\vec{a}_{ij}$ denote the author metadata associated with article j in the matched pair i. Such a matched-pair design enables us to test the following sharp null hypothesis:

$$H'_{0,\text{sharp}}: \quad Y_{ij}(\vec{a}_{i1}) = Y_{ij}(\vec{a}_{i2}), \quad \forall i = 1,\ldots,I, j = 1, 2. \tag{2}$$

We note that $H_{0,\text{sharp}}$ in (1) implies $H'_{0,\text{sharp}}$ in (2), so rejecting $H'_{0,\text{sharp}}$ would then provide evidence against $H_{0,\text{sharp}}$. Such a design is termed "near-far" design in the literature [32,33] and has been used in a number of empirical studies [34–36].

4 Data analysis: study design and outcome analysis

4.1 A first matched comparison: design M1

We restricted our attention to 5,313 borderline articles that were peer-reviewed by three or four reviewers and received an average rating between 5 and 7. We first considered a study design M1 where each matched pair consisted of one article whose authors' average institution ranking was among the top 30% of these 5,313 submissions and the other article whose authors' average institution ranking was among the bottom 70%. Columns 2 and 3 of Table 1 summarize the characteristics, including structural features and derived higher-level features, of 1,585 top-30% articles and those of the other 3,728 articles. As one closely examines these two columns, a number of features, including submission year, number of figures, complexity score as judged by the language model, keyword, and topic, differ systematically among these submissions. Matching helps remove most of the overt bias: in the matched comparison group (which is a subset of size $n = 1{,}585$ from the reservoir of 3,728 articles), standardized mean differences of all but one covariates are less than 0.1, or one-tenth of one pooled standard deviation. In fact, design M1 required near-exact matching on important covariates such as reviewers' ratings and year of submission and achieved near-fine balance for categorical variables such as topic cluster and primary keyword [37]. Algorithms used to construct the matched design M1 will be described in detail in Section 4.3. Panel c in Figure 2 assesses how different two articles in each matched pair are in their authors' average institution rankings (Figure A1 in Appendix for similar plots in the four-reviewer stratum). The average, within-matched-pair difference in authors' average institution ranking is 74.3 among 983 matched pairs in the three-reviewer stratum (median: 49.6; interquartile range: 26.6–96.7) and 99.6

Table 1: Characteristics of articles before and after matching in the design M1

	Bottom 70% (n = 3,728)	Top 30% (n = 1,585)	SMD* before	M1 (n = 1,585)	SMD after
Conference and reviewer					
Year of submission (%)					
2017	109 (2.9)	129 (8.1)	0.23	92 (5.8)	0.10
2018	299 (8.0)	221 (13.9)	0.19	201 (12.7)	0.04
2019	520 (13.9)	303 (19.1)	0.14	304 (19.2)	<0.01
2020	548 (14.7)	231 (14.6)	<0.01	234 (14.8)	<0.01
2021	1200 (32.2)	388 (24.5)	0.17	412 (26.0)	0.03
2022	1052 (28.2)	313 (19.7)	0.20	342 (21.6)	0.05
Reviewer ratings					
Reviewer I	6.99 (0.92)	6.99 (0.91)	<0.01	6.97 (0.90)	0.02
Reviewer II	6.12 (0.80)	6.12 (0.79)	<0.01	6.12 (0.78)	<0.01
Reviewer III	5.21 (1.08)	5.18 (1.09)	0.03	5.19 (1.10)	0.01
Reviewer IV**	4.71 (1.06)	4.74 (1.11)	0.03	4.74 (1.09)	<0.01
Reviewer sentiment					
Reviewer I	0.75 (0.10)	0.75 (0.11)	0.03	0.75 (0.11)	0.03
Reviewer II	0.64 (0.09)	0.64 (0.09)	0.02	0.64 (0.09)	0.04
Reviewer III	0.55 (0.10)	0.54 (0.10)	0.08	0.54 (0.10)	0.01
Reviewer IV§	0.49 (0.09)	0.50 (0.09)	0.06	0.50 (0.09)	0.01
Article metadata					
No. author	4.21 (1.67)	4.17 (1.69)	0.02	4.13 (1.67)	0.03
No. figure	13.71 (7.26)	12.55 (7.55)	0.16	12.42 (6.60)	0.02
No. reference	42.53 (16.59)	42.19 (16.93)	0.02	40.98 (14.94)	0.07
No. section	19.94 (7.16)	19.96 (7.11)	<0.01	19.74 (6.90)	0.03
Complexity, topics, and keywords					
Complexity	0.84 (0.03)	0.85 (0.03)	0.29	0.85 (0.03)	0.06
Topic cluster† (%)					
RL/meta learning/robustness	367 (9.8)	113 (7.1)	0.10	113 (7.1)	0
RL/CV/robustness	298 (8.0)	80 (5.0)	0.12	80 (5.0)	0
DL/GM/CNN	345 (9.3)	147 (9.3)	0	147 (9.3)	0
DL/RNN/GNN	365 (9.8)	133 (8.4)	0.05	133 (8.4)	0
DL/optimization/generalization	399 (10.7)	126 (7.9)	0.10	126 (7.9)	0
DL/robustness/adversarial examples	445 (11.9)	270 (17.0)	0.15	270 (17.0)	0
DL/RL/unsupervised learning/GM	319 (8.6)	143 (9.0)	0.01	143 (9.0)	0
DL/multi-agent or model-based RL/IL	475 (12.7)	209 (13.2)	0.02	209 (13.2)	0
DL/federated or distributed learning	370 (9.9)	260 (16.4)	0.19	260 (16.4)	0
GM/GAN/VAE	345 (9.3)	104 (6.6)	0.10	104 (6.6)	0
Primary keyword (%)					
NA	950 (25.5)	347 (21.9)	0.09	368 (23.2)	0.03
Other	794 (21.3)	292 (18.4)	0.07	312 (19.7)	0.03
Deep learning	393 (10.5)	242 (15.3)	0.14	238 (15.0)	0.01
Reinforcement learning	290 (7.8)	183 (11.5)	0.13	181 (11.4)	<0.01
Graph neural networks	145 (3.9)	41 (2.6)	0.07	39 (2.5)	<0.01
Representation learning	109 (2.9)	40 (2.5)	0.03	39 (2.5)	<0.01
Generative models	89 (2.4)	35 (2.2)	0.01	38 (2.4)	0.01
Meta-learning	79 (2.1)	34 (2.1)	0	33 (2.1)	<0.01
Self-supervised learning	72 (1.9)	23 (1.5)	0.03	24 (1.5)	<0.01
Unsupervised learning	70 (1.9)	43 (2.7)	0.05	32 (2.0)	0.05
Neural networks	62 (1.7)	30 (1.9)	0.02	25 (1.6)	0.02
Generative adversarial networks	56 (1.5)	14 (0.9)	0.06	18 (1.1)	0.02
Optimization	43 (1.2)	26 (1.6)	0.03	26 (1.6)	0
Variational inference	39 (1.0)	8 (0.5)	0.06	11 (0.7)	0.02
Transformer	37 (1.0)	20 (1.3)	0.03	15 (0.9)	0.04
Generalization	36 (1.0)	32 (2.0)	0.08	22 (1.4)	0.05

(Continued)

Table 1: *Continued*

	Bottom 70% (n = 3,728)	Top 30% (n = 1,585)	SMD* before	M1 (n = 1,585)	SMD after
Decision					
Acceptance (%)	1,928 (51.7)	811 (51.2)		851 (53.7)	

* SMD: Standardized mean differences.
§ Reviewer IV's rating and sentiment results are derived from articles within the stratum of four reviewers.
† RL: Reinforcement learning; GM: Generative models; CV: Computer vision; CNN: Convolutional neural nets; RNN: Recurrent neural nets; GNN: Graph neural nets; IL: Imitation learning; GAN: Generative adversarial nets; VAE: Variational auto-encoder. Note that the description is not exhaustive.

among 602 matched pairs in the four-reviewer stratum (median: 63.5; interquartile range: 28.5–135.0). On the other hand, the average, within-matched-pair difference in the highest ranked institution among co-authors is 43.8 among 983 matched pairs in the three-reviewer stratum (median: 20.5; interquartile range: 6.0–50.3) and 55.6 among 602 matched pairs in the four-reviewer stratum (median: 20.0; interquartile range: 6.0–56.5). We concluded that there was a sizable difference in author metadata between two articles in each matched pair.

To further demonstrate two groups are well comparable, Panel a of Figure 2 displays the distribution of the estimated "propensity score," defined as the probability that authors' average ranking of an article was among top 30%, in each of the following three groups: top-30% articles (red), bottom-70% articles (blue), and matched comparison articles (yellow), all in the three-reviewer stratum. Similar plots for articles with four reviewers can be found in Appendix. As is evident from the figure, the propensity score distribution of the matched comparison articles is more similar to that of the top-30% articles. Panel b of Figure 2 further plots the cosine similarity calculated from the raw textual data of each article. It is also evident that two articles

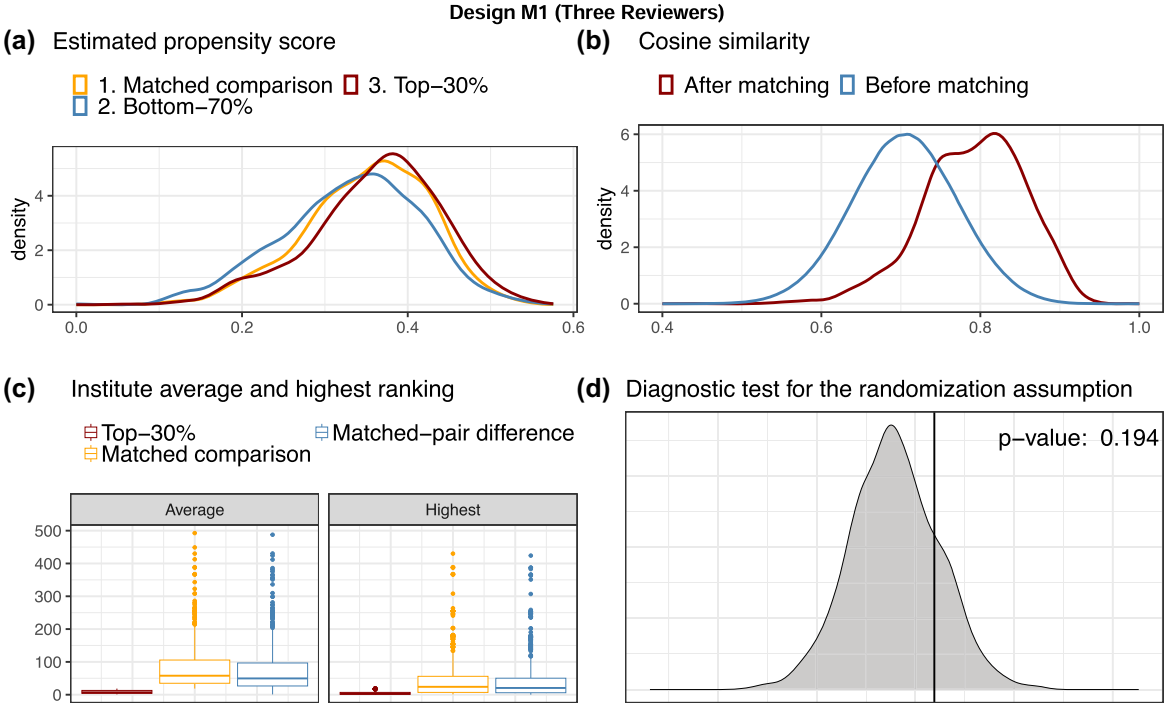

Figure 2: Diagnostics of design M1. Panel a: estimated propensity score top-30% articles, bottom-70% articles and matched comparison articles; Panel b: distribution of between-article cosine similarity before (median = 0.71) and after matching (median = 0.80); Panel c: boxplots of authors' average (highest) institution rankings and matched-pair differences in authors' average (highest) institution rankings; Panel d: permutation distribution and the observed test statistic of the classification permutation test (CPT).

Table 2: Contingency table, p-values (exact, two-sided McNemar's test), and odds ratio [39] of 1,585 matched pairs in the design M1 and 1,051 matched pairs in a strengthened design M2 and the associated subgroup analyses among matched pairs where both papers had a score below 6 and those where both had a score above 6 (inclusive)

		Comparison articles			Odds ratio
		Accepted	Rejected	P-value	95% CI
		Panel a: Design M1			
Overall					
Top-30% articles	Accepted	633	178	0.050	0.82
	Rejected	218	556		[0.67, 1.00]
Subgroup: Avg. < 6					
Top-30% articles	Accepted	41	60	<0.001	0.57
	Rejected	106	482		[0.41, 0.78]
Subgroup: Avg. ≥ 6					
Top-30% articles	Accepted	580	93	0.824	1.04
	Rejected	89	66		[0.77, 1.41]
		Panel b: Design M2			
Overall					
Top-20% articles	Accepted	443	115	0.149	0.83
	Rejected	139	354		[0.64, 1.07]
Subgroup: Avg. < 6					
Top-20% articles	Accepted	37	39	0.018	0.61
	Rejected	64	305		[0.40, 0.92]
Subgroup: Avg. ≥ 6					
Top-20% articles	Accepted	399	59	0.426	0.86
	Rejected	69	42		[0.59, 1.23]

Defining and interpreting the odds ratio and its confidence interval requires an additional "stability" assumption discussed in Section 3.2.

in the same matched pair now have improved cosine similarity compared to that from two randomly drawn articles prior to matching. Our designed matched comparison M1 appears to be well balanced in many observed covariates and resembles a hypothetical RCT where we randomly assign author metadata.

The question remains as to whether the balance is sufficiently good compared to an authentic RCT and could justify randomization inference. We conducted a formal diagnostic test using Gagnon–Bartsch and Shem–Tov's [38] CPT based on random forests to test the randomization assumption for the matched cohort. The randomization assumption cannot be rejected in either the three-reviewer or four-reviewer stratum (p-value = 0.194 and 0.641, respectively) (see the null distribution and test statistic in Panel d of Figure 2 and Figure A1 in Appendix).

Panel a of Table 2 summarizes the outcomes of 1,585 matched pairs of two articles. We tested Fisher's sharp null hypothesis $H'_{0,\text{sharp}}$ reviewed and discussed in Section 3.4 using a two-sided, exact McNemar's test [39] and obtained a p-value of 0.050, which suggested some weak evidence that authorship metadata was associated with AC's final decisions. We further examined the effect heterogeneity by repeating the analysis among matched pairs where both articles had an average score below 6 and matched pairs where both articles had an average score greater than or equal to 6. Interestingly, among the subgroup of below-6 borderline articles, we had strong evidence that author metadata was associated with area chairs' final decisions (p-value <0.001). Under an additional stability assumption stating that the potential acceptance status $Y(\vec{a})$ depended on author metadata $\vec{a}$ only via authors' average institution ranking and remained unchanged when the average ranking is among the top-30% or among the bottom-70%, we estimated the odds ratio (OR) to be 0.57 (95% CI: [0.41, 0.78]), providing evidence that below-6 borderline articles from top-30% institutions were *less* favored by area chairs compared to their counterparts in the comparison group. On the other hand, among borderline articles with an average score greater than or equal to 6, we had no evidence that author metadata was associated with area chairs' final decisions (OR = 1.04; 95% CI: [0.77, 1.41]; p-value = 0.824).

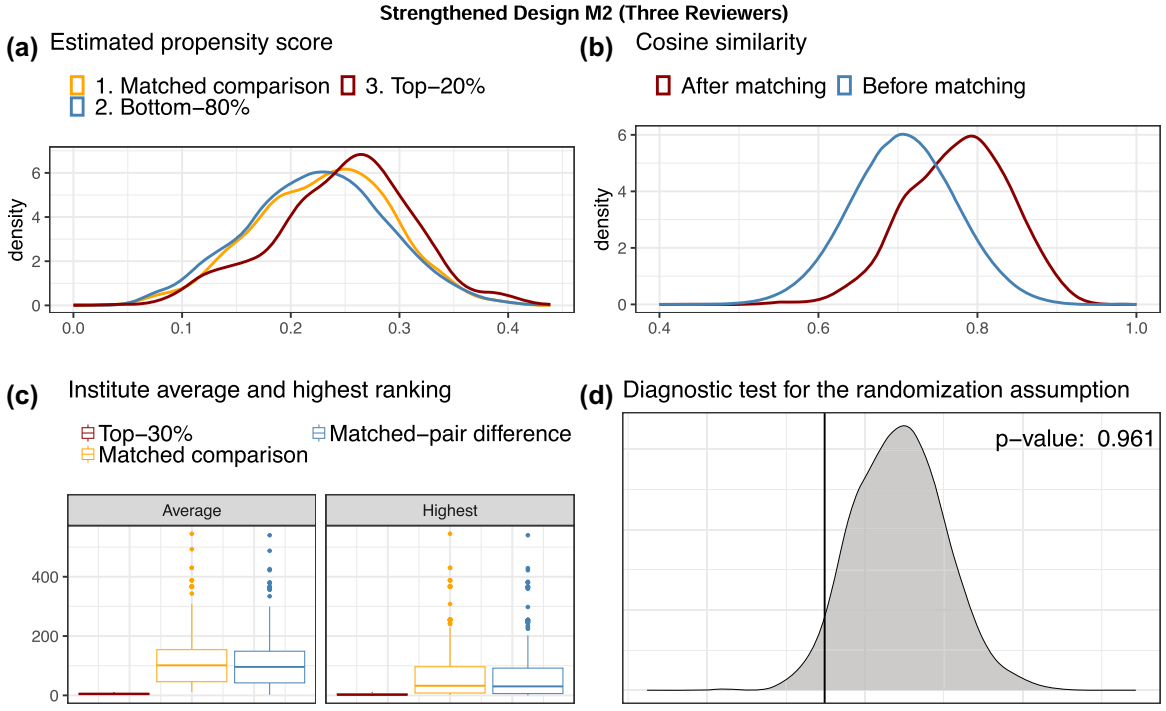

Figure 3: Diagnostics of design M2. Panel a: estimated propensity score top-30% articles, bottom-70% articles and matched comparison articles; Panel b: distribution of between-article cosine similarity before (median = 0.71) and after matching (median = 0.78). Panel c: boxplots of authors' average (highest) institution rankings and matched-pair differences in authors' average (highest) institution rankings. Panel d: permutation distribution and the observed test statistic of the CPT.

Finally, we conducted a sensitivity analysis where we excluded matched pairs with at least one article arxived before the end of the peer-review process. Approximately 15% of all submissions in our dataset were matched to their arxived versions. By excluding these articles, reviewers' ratings and comments were more likely to be a faithful reflection of the true quality of the article. Because the proportion of arxived papers was relatively small, results remained similar (below-6 submissions: OR = 0.61; 95% CI: [0.41, 0.89]; p-value = 0.001; above-6 submissions: OR = 1.16; 95% CI: [0.82, 1.66]; p-value = 0.437).

Our analysis seems to defy the common wisdom that there is a "status bias" favoring higher-profile authors. It is then natural to ask, if author metadata were even more drastically different, shall we then see some evidence for status bias that would better align with previous findings? This inquiry led to a second, strengthened design M2.

4.2 Strengthened design M2

Baiocchi et al. [32] first considered "strengthening" an instrumental variable (the differential distance) in an analysis of the effect of low-level versus high-level neonatal intensive care units (NICUs) on mortality rate. In their analysis, Baiocchi et al. [32] compared mothers who lived near to and far from a high-level NICU and "strengthened" the comparison by restricting their analysis to a smaller subset of comparable mothers who lived very near to and very far from a high-level NICU (see also Zhang et al. [40] and Chen et al. [41] for some related recent development). We adopted this idea here and constructed a strengthened design M2 where one article in each matched pair is a top-20% article (as opposed to a top-30% article in design M1) and the other matched comparison article was from the reservoir of bottom-80% articles. We also added a "dose-caliper" in the statistical matching algorithm to further separate the average ranking within each matched pair (see Section 4.3 for details).

Table 3: Characteristics of articles before and after matching in a strengthened design M2

	Bottom 80% (*n* = 4,262)	Top 20% (*n* = 1,051)	SMD before	M2 (*n* = 1,051)	SMD after
Conference and reviewer					
Year of submission (%)					
2017	144 (3.4)	94 (8.9)	0.23	77 (7.3)	0.07
2018	380 (8.9)	140 (13.3)	0.14	126 (12.0)	0.04
2019	614 (14.4)	209 (19.9)	0.15	201 (19.1)	0.02
2020	621 (14.6)	158 (15.0)	0.01	157 (14.9)	<0.01
2021	1343 (31.5)	245 (23.3)	0.19	272 (25.9)	0.06
2022	1160 (27.2)	205 (19.5)	0.18	218 (20.7)	0.03
Reviewer ratings					
Reviewer I	6.98 (0.92)	7.04 (0.91)	0.06	7.01 (0.89)	0.02
Reviewer II	6.12 (0.79)	6.13 (0.79)	0.02	6.11 (0.78)	0.03
Reviewer III	5.21 (1.07)	5.16 (1.11)	0.04	5.17 (1.09)	<0.01
Reviewer IV*	4.71 (1.06)	4.77 (1.12)	0.05	4.74 (1.13)	0.02
Reviewer sentiment					
Reviewer I	0.75 (0.10)	0.76 (0.11)	0.05	0.75 (0.10)	0.10
Reviewer II	0.64 (0.09)	0.64 (0.09)	<0.01	0.64 (0.09)	0.08
Reviewer III	0.55 (0.10)	0.54 (0.10)	0.06	0.54 (0.10)	0.03
Reviewer IV*	0.49 (0.09)	0.50 (0.09)	0.12	393	0.06
Article metadata					
No. author	4.21 (1.67)	4.17 (1.72)	0.02	4.19 (1.66)	0.01
No. figure	13.60 (7.41)	12.41 (7.10)	0.16	12.32 (6.45)	0.01
No. reference	42.56 (16.61)	41.91 (17.02)	0.04	40.50 (15.12)	0.08
No. section	19.94 (7.16)	19.96 (7.06)	<0.01	19.54 (6.94)	0.06
Complexity, topics, and keywords					
Complexity	0.84 (0.03)	0.85 (0.03)	0.31	0.85 (0.03)	0.15
Topic cluster† (%)					
RL/meta learning/robustness	404 (9.5)	76 (7.2)	0.08	76 (7.2)	0
RL/CV/robustness	328 (7.7)	50 (4.8)	0.12	50 (4.8)	0
DL/GM/CNN	393 (9.2)	99 (9.4)	<0.01	99 (9.4)	0
DL/RNN/GNN	408 (9.6)	90 (8.6)	0.04	90 (8.6)	0
DL/optimization/generalization	439 (10.3)	86 (8.2)	0.07	86 (8.2)	0
DL/robustness/adversarial examples	551 (12.9)	164 (15.6)	0.08	164 (15.6)	0
DL/RL/unsupervised learning/GM	365 (8.6)	97 (9.2)	0.02	97 (9.2)	0
DL/multi-agent or model-based RL/IL	545 (12.8)	139 (13.2)	0.01	139 (13.2)	0
DL/federated or distributed learning	435 (10.2)	195 (18.6)	0.24	195 (18.6)	0
GM/GAN/VAE	394 (9.2)	55 (5.2)	0.16	55 (5.2)	0
Primary keyword (%)					
NA	1084 (25.4)	213 (20.3)	0.12	220 (20.9)	0.01
Other	891 (20.9)	195 (18.6)	0.06	195 (18.6)	0
Deep learning	460 (10.8)	175 (16.7)	0.17	178 (16.9)	<0.01
Reinforcement learning	353 (8.3)	120 (11.4)	0.10	122 (11.6)	<0.01
Graph neural networks	162 (3.8)	24 (2.3)	0.09	24 (2.3)	0
Representation learning	127 (3.0)	22 (2.1)	0.06	22 (2.1)	0
Generative models	99 (2.3)	25 (2.4)	<0.01	26 (2.5)	<0.01
Meta-learning	89 (2.1)	24 (2.3)	0.01	23 (2.2)	<0.01
Self-supervised learning	80 (1.9)	15 (1.4)	0.04	14 (1.3)	<0.01
Unsupervised learning	79 (1.9)	34 (3.2)	0.08	26 (2.5)	0.04
Neural networks	73 (1.7)	19 (1.8)	<0.01	17 (1.6)	0.02
Generative adversarial networks	62 (1.5)	8 (0.8)	0.07	10 (1.0)	0.02
Generalization	48 (1.1)	20 (1.9)	0.07	16 (1.5)	0.03
Optimization	47 (1.1)	22 (2.1)	0.08	22 (2.1)	0
Variational inference	44 (1.0)	3 (0.3)	0.09	7 (0.7)	0.05
Decision					
Acceptance (%)	2,181 (51.2)	558 (53.1)		582 (55.4)	

* Reviewer IV's rating and sentiment results are derived from articles within the stratum of four reviewers.

† RL: reinforcement learning; GM: generative models; CV: computer vision; CNN: convolutional neural nets; RNN: recurrent neural nets; GNN: graph neural nets; IL: imitation learning; GAN: generative adversarial nets; VAE: variational auto-encoder. Note that the description is not exhaustive.

A total of 1,051 matched pairs were formed. Panel c of Figure 3 summarizes the within-matched-pair difference in authors' average institution ranking across 658 matched pairs in the three-reviewer stratum of the strengthened design M2. In this strengthened design, the matched-pair-difference in institution ranking (averaged over all co-authors) now increases from 74.3 (as in the design M1) to 108.8, and the difference in the ranking of the highest ranked co-author increases from 43.8 to 64.4. Importantly, the cohort of top-20% articles and their matched comparison group are still comparable in all baseline covariates, as summarized in Table 3. Similar to the design M1, we cannot reject the randomization assumption based on the CPT (p-value = 0.961).

Panel b of Table 2 summarizes the outcomes of 1,051 matched pairs of two articles in the strengthened design M2 and associated subgroup analyses. We observed results similar to those derived from the design M1, although the estimates were slightly less precise as a result of a smaller sample size (1,051 in M2 versus 1,585 in M1). Consistent results across two study designs help reinforce the conclusion that we did not find evidence supporting a "status bias" favoring authors from high-ranking institutions in this cohort of borderline articles. Quite contrary, below-6 borderline articles from top institutions appeared to be less favored by area chairs.

4.3 Matching algorithm: matching one sample according to multiple criteria

The matched sample M1 displayed in Table 1 was constructed using an efficient, network-flow-based optimization algorithm built upon a tripartite network [42] as opposed to a traditional, bipartite network [43]. Compared to a bipartite network, a tripartite network structure helps separate two tasks in the design of an observational study: (1) constructing closely matched pairs and (2) constructing a well-balanced matched sample. A detailed account of the algorithm can be found in the study of Zhang et al. [42]; in the following, we described how we designed the cost in the tripartite network and achieved the features of M1 and M2 described in Sections 4.1 and 4.2.

Figure 4 illustrates the basic tripartite network structure with three units, $\{\gamma_1, \gamma_2, \gamma_3\}$, from the top-30% articles and five units, $\{\tau_1, \ldots, \tau_5\}$, from the reservoir of bottom-70% articles. The goal of tripartite-graph-based statistical matching algorithm is to select a subset of units from $\{\tau_1, \ldots, \tau_5\}$ so that this selected subset resembles $\{\gamma_1, \gamma_2, \gamma_3\}$ in some aspects quantified by δ_{γ_i,τ_j} and possibly some other aspects quantified by $\Delta_{\overline{\gamma}_i,\overline{\tau}_j}$. This can be efficiently achieved by running a network-flow-based algorithm, where a feasible flow of magnitude 3 emanates from the source node ξ, subsequently travels through the layer $\{\gamma_1, \ldots, \gamma_3\}$, layer $\{\tau_1, \ldots, \tau_5\}$, layer $\{\overline{\tau}_1, \ldots, \overline{\tau}_5\}$, layer $\{\overline{\gamma}_1, \ldots, \overline{\gamma}_3\}$, and finally arrives at the sink $\overline{\xi}$. The blue edges in Figure 4 illustrate one feasible flow, which picks τ_1, τ_2, τ_5 and pairs τ_1 to γ_1, τ_2 to γ_2, and τ_5 to γ_3.

Each feasible flow is associated with a total cost [42]. This total cost consists of two important pieces. The *cost* δ_{γ_i,τ_j} associated with each edge e connecting γ_i and τ_j in the left part of the network encodes criteria

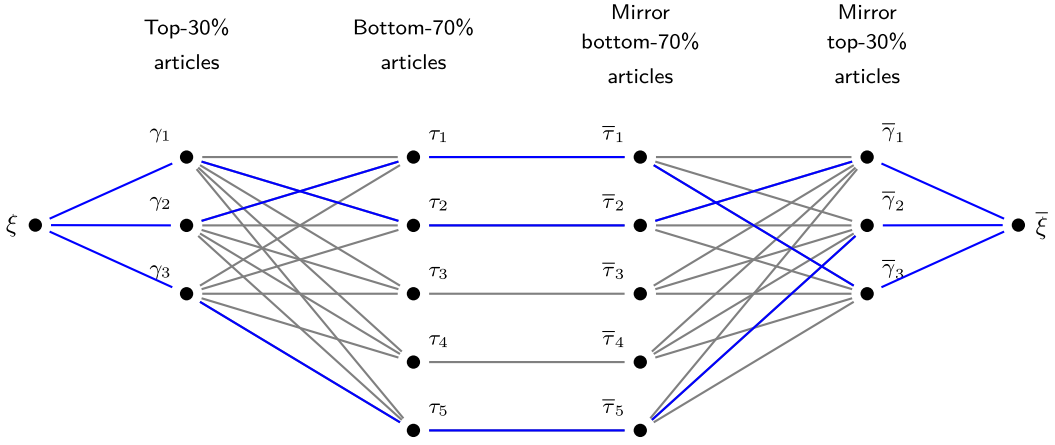

Figure 4: An illustrative plot of a tripartite network consisting of three treated units and five candidate control units. The blue edges correspond to one feasible flow that selects $\{\tau_1, \tau_2, \tau_5\}$ and pairs τ_1 to γ_2, τ_2 to γ_1, and τ_5 to γ_3.

for closely matched pairs. To construct the matched sample M1, we let δ_{γ_i,τ_j} be the cosine similarity derived from the SPECTER embeddings of article γ_i and τ_j. We then enforce near-exact matching on submission year and reviewers' ratings by adding a large penalty to δ_{γ_i,τ_j} if articles γ_i and τ_j were submitted to the ICLR conference in different calendar years or did not receive same ratings. As a result, two articles in each matched pair were exactly or near-exactly matched on reviewers' ratings. For instance, a top-30% article γ_i submitted at 2017 and peer-reviewed by three reviewers and received ratings of 6, 5 and 5 (from highest to lowest) was matched to a 2017-submitted, (6, 5, 5)-rated bottom-70% article τ_j. This conscious design of δ_{γ_i,τ_j} helped achieve improved within-matched-pair cosine similarity and near-exact match on year of submission and reviewers' ratings in M1.

The cost $\Delta_{\overline{\gamma}_i,\overline{\tau}_j}$ associated with edge connecting $\overline{\gamma}_i$ and $\overline{\tau}_j$ in the right part of the network encodes criteria for good overall balance when groups are viewed as a whole. To construct M1, we estimated the propensity score based on article metadata and set $\Delta_{\overline{\gamma}_i,\overline{\tau}_j}$ to be the difference in the estimated propensity score to minimize earth-mover's distance between the propensity score distributions of the top-30% articles and their matched comparison articles. The "balancing" property of the propensity score [26] then helped balance the covariates used to estimate it. One limitation of the propensity score is that its stochastic balancing property often does not suffice when balancing nominal variables with many categories; in these scenarios, a design technique known as *fine balance* is often used in conjunction with propensity score matching [11,37]. In short, fine balance is a technique that forces the frequency of one or more nominal variables to be identical or as close as possible in two groups. We finely balanced the topic clusters and keywords by adding a large penalty to $\Delta_{\overline{\gamma}_i,\overline{\tau}_j}$ when γ_i and τ_j differed in the topic cluster or keyword. Finally, matched pairs in M1 were obtained as a result of solving the minimum-cost flow optimization problem associated with this tripartite network. We conducted matching in the stratum of articles with three reviewers and four reviewers, respectively, because articles in the four-reviewer stratum have two additional covariates: a fourth reviewer rating and a fourth reviewer sentiment.

Our construction of the design M2 was analogous to that of M1, except that we further added a "dose-caliper," defined as a large penalty when two articles γ_i and τ_j differ in their authors' average institution rankings by less than a pre-specified caliper size, to the cost δ_{γ_i,τ_j}. The average within-matched-pair difference in authors' average rankings is 74.3 in M1; hence, we set the caliper size to be 80 when constructing M2. In this way, the within-matched-pair difference in authors' average ranking was as large as 108.8 in the design M2, representing a meaningful improvement over that in M1.

5 Conclusion: interpretation of our findings; limitations

In this article, we proposed a quasi-experimental design approach to studying the association between author metadata and area chairs' decisions. We did *not* find evidence supporting a *status bias*, i.e., area chairs' decisions systematically favored authors from high-ranking institutions, when comparing two cohorts of borderline articles with near-identical reviewers' ratings, sentiment, topics, primary keywords, and article metadata. Under an additional stability assumption, we found that articles from high-ranking institutions had a lower acceptance rate. The result was consistent among articles from top-30% institutions (odds ratio = 0.82) and top-20% institutions (odds ratio = 0.83) and most pronounced among the subgroup of below-6 submissions (p-value <0.001 and 0.018 in two designs, respectively).

Our results need to be interpreted under an appropriate context. First, like all retrospective, observational studies, although we formulated the question under a rigorous causal framework, we cannot be certain that our analysis was immune from any unmeasured confounding bias. For instance, the marginally higher acceptance rate of articles in the matched comparison groups (i.e., articles from lower-ranking institutions) could be easily explained away if these articles, despite having near-identical reviewers' ratings as their counterparts in the top-30% or top-20% groups, were in fact superior in their novelty and significance and area chairs made decisions based on these attributes rather than author metadata.

Second, any interpretation of our results should be restricted to our designed matched sample and should *not* be generalized to other contexts. In particular, we only focused on area chairs' decision on *borderline*

articles. As Greiner and Rubin [7, Section III] articulated in great detail, a study unit may interact with multiple agents of an overall system and we have more than one choice of decider to study. In our case study, an ICLR article has interacted with at least two types of deciders, a group of reviewers and an area chair. We explicitly focused on the interaction between a peer-reviewed article and an area chair. This deliberate choice allowed us to control for some valuable pre-treatment variables, including reviewers' ratings and sentiment, that are good proxies for articles' innate quality; however, by choosing to focus on this interaction that happened after the interaction between an article and multiple reviewers, we forwent the opportunity to detect any status bias, in either direction, in any earlier interaction including the peer-review process that could have affected the values of pre-treatment variables in our analysis [7]. Although the peer-review process of ICLR is in principle double-blind, it is conceivable that articles' author metadata could be leaked (e.g., when articles were posted on the pre-print platform or when the metadata could be inferred from articles' cited related works and writing style) during the peer-review process, and reviewers' ratings could be biased in favor of more (or less) established authors. It is of great interest to further study any perception bias during the interaction between articles and their reviewers; however, a key complication facing such an analysis is that articles may not be comparable without having a relatively objective judgment or rating (e.g., reviewers' ratings in our analysis of area chairs' decision).

With all these important caveats and limitations in mind, we found our analysis a solid contribution to the social science literature on status bias. Our analysis also helps clarify many important causal inference concepts and misconceptions when study units are textual data, including (1) the importance of shifting focus from an attribute to the perception of it, (2) the importance of articulating the timing of the treatment and hence what constitutes pre-treatment variables, (3) Fisher's sharp null hypothesis as a relevant causal null hypothesis in the context of fairness, and (4) Rubin's [28] stability assumption often implicitly assumed but overlooked, all within a concrete case study.

Acknowledgement: The authors are grateful for the reviewer's valuable comments that improved the manuscript.

Funding information: The authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript. CC, JZ, and BZ designed the study, prepared the data and analyses, and prepared the manuscript. TY and DR provided insight into the design and analysis of the current study.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: The datasets generated during and/or analyzed during the current study are available in the repository: https://zjiayao.github.io/iclr_database/.

References

[1] Cortes C, Lawrence ND. Inconsistency in conference peer-review: Revisiting the 2014 NeurIPS experiment. 2021. https://arxiv.org/abs/2109.09774.

[2] McGillivray B, De Ranieri E. Uptake and outcome of manuscripts in Nature journals by review model and author characteristics. Res Integrity Peer Rev. 2018;3(1):5.

[3] Tomkins A, Zhang M, Heavlin WD. Reviewer bias in single-versus double-blind peer-review. Proc Nat Acad Sci. 2017;114(48):12708–13.

[4] Sun M, Barry Danfa J, Teplitskiy M. Does double-blind peer-review reduce bias? Evidence from a top computer science conference. J Assoc Inform Sci Tech. 2022;73(6):811–9.

[5] Smirnova I, Romero DM, Teplitskiy M. The bias-reducing effect of voluntary anonymization of authors' identities: Evidence from peer review (January 27, 2023). 2022. SSRN 4190623.
[6] Bertrand M, Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Amer Econ Rev. 2004;94(4):991–1013.
[7] Greiner DJ, Rubin DB. Causal effects of perceived immutable characteristics. Rev Econ Stat. 2011;93(3):775–85.
[8] Huber J, Inoua S, Kerschbamer R, König-Kersting C, Palan S, Smith VL. Nobel and novice: Author prominence affects peer-review. Proc Nat Acad Sci. 2022;119(41):e2205779119.
[9] Nielsen MW, Baker CF, Brady E, Petersen MB, Andersen JP. Weak evidence of country-and institution-related status bias in the peer-review of abstracts. Elife. 2021;10:e64561.
[10] Rosenbaum PR. Observational studies. New York: Springer; 2002.
[11] Rosenbaum PR. Design of Observational Studies. New York: Springer; 2010.
[12] Neyman JS. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Ann Agricult Sci. 1923;10:1–51.
[13] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol. 1974;66(5):688.
[14] Zhang J, Zhang H, Deng Z, Roth D. Investigating fairness disparities in peer review: A language model enhanced approach. 2022. https://arxiv.org/abs/2211.06398.
[15] Tsai CT, Mayhew S, Roth D. Cross-lingual named entity recognition via wikification. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics; 2016. p. 219–28. https://aclanthology.org/K16-1022.
[16] Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. SPECTER: document-level representation learning using citation-informed transformers; 2020. https://arxiv.org/abs/2004.07180.
[17] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT; 2019. p. 4171–86.
[18] van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Machine Learn Res. 2008;9(86):2579–605. http://jmlr.org/papers/v9/vandermaaten08a.html.
[19] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. CoRR. 2019;1907.11692. http://arxiv.org/abs/1907.11692.
[20] Rosenthal S, Farra N, Nakov P. SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017); 2017. p. 502–18.
[21] Cochran WG, Chambers SP. The planning of observational studies of human populations. J R Stat Soc Ser A (General). 1965;128(2):234–66.
[22] Rubin DB. Matching to remove bias in observational studies. Biometrics. 1973;29:159–83.
[23] Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. J Amer Stat Assoc. 1979;74(366):318–28.
[24] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Politic Anal. 2007;15(3):199–236.
[25] Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010;25(1):1–21.
[26] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
[27] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Amer Statist. 1985;39(1):33–8.
[28] Rubin D. Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. J Amer Stat Assoc. 1980;75:591–3.
[29] Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. J Amer Stat Assoc. 2005;100(469):322–31.
[30] Zhang B, Zhang J. Some reflections on drawing causal inference using textual data: Parallels between human subjects and organized texts. In: Schölkopf B, Uhler C, Zhang K, editors. Proceedings of the First Conference on Causal Learning and Reasoning. vol. 177 of Proceedings of Machine Learning Research. PMLR; 2022. p. 1026–36. https://proceedings.mlr.press/v177/zhang22b.html.
[31] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Int Med. 2017;167(4):268–74.
[32] Baiocchi M, Small DS, Lorch S, Rosenbaum PR. Building a stronger instrument in an observational study of perinatal care for premature infants. J Amer Stat Assoc. 2010;105(492):1285–96.
[33] Baiocchi M, Small DS, Yang L, Polsky D, Groeneveld PW. Near/far matching: a study design approach to instrumental variables. Health Services Outcomes Res Methodol. 2012;12(4):237–53.
[34] Lorch SA, Baiocchi M, Ahlberg CE, Small DS. The differential impact of delivery hospital on the outcomes of premature infants. Pediatrics. 2012;130(2):270–8.
[35] Neuman MD, Rosenbaum PR, Ludwig JM, Zubizarreta JR, Silber JH. Anesthesia technique, mortality, and length of stay after hip fracture surgery. JAMA. 2014;311(24):2508–17.
[36] MacKay EJ, Zhang B, Heng S, Ye T, Neuman MD, Augoustides JG, et al. Association between transesophageal echocardiography and clinical outcomes after coronary artery bypass graft surgery. J Amer Soc Echocardiography. 2021;34(6):571–81. https://www.sciencedirect.com/science/article/pii/S0894731721000298.

[37] Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. J Amer Stat Assoc. 2007;102(477):75–83.
[38] Gagnon-Bartsch J, Shem-Tov Y. The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. Ann Appl Stat. 2019;13(3):1464–83.
[39] Fay MP. Two-sided exact tests and matching confidence intervals for discrete data. The R Journal. 2010;2(1):53–58. doi: 10.32614/RJ-2010-008.
[40] Zhang B, Mackay EJ, Baiocchi M. Statistical matching and subclassification with a continuous dose: Characterization, algorithm, and application to a health outcomes study. Ann Appl Stat. 2023;17(1):454–75.
[41] Chen Z, Cho MH, Zhang B. Manipulating a continuous instrumental variable in an observational study of premature babies: Algorithm, partial identification bounds, and inference under randomization and biased randomization assumptions. 2024. arXiv:240417734.
[42] Zhang B, Small D, Lasater K, McHugh M, Silber J, Rosenbaum P. Matching one sample according to two criteria in observational studies. J Amer Stat Assoc. 2021;(just-accepted):1–34.
[43] Rosenbaum PR. Optimal matching for observational studies. J Amer Stat Assoc. 1989;84(408):1024–32.

Appendix

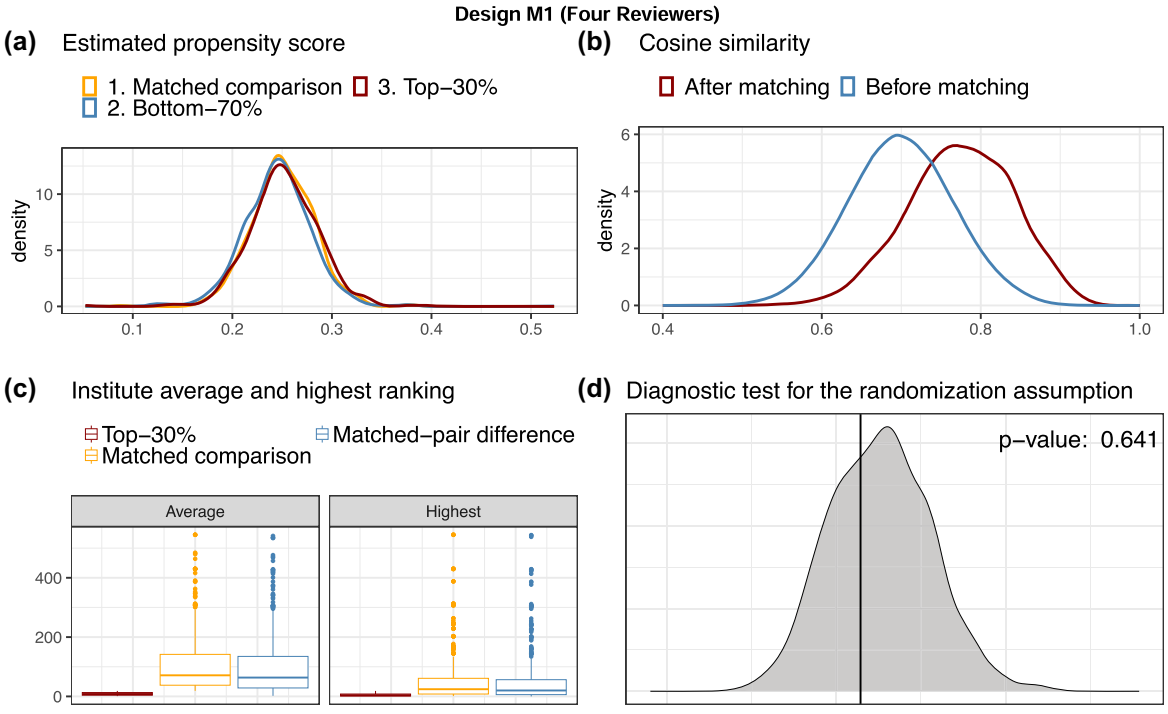

Figure A1: Diagnostics of design M1. Panel a: estimated propensity score top-30% articles, bottom-70% articles, and matched comparison articles; Panel b: distribution of between-article cosine similarity before (median = 0.70) and after matching (median = 0.78); Panel c: boxplots of authors' average (highest) institution rankings and matched-pair differences in authors' average (highest) institution rankings; Panel d: permutation distribution and the observed test statistic of the CPT.

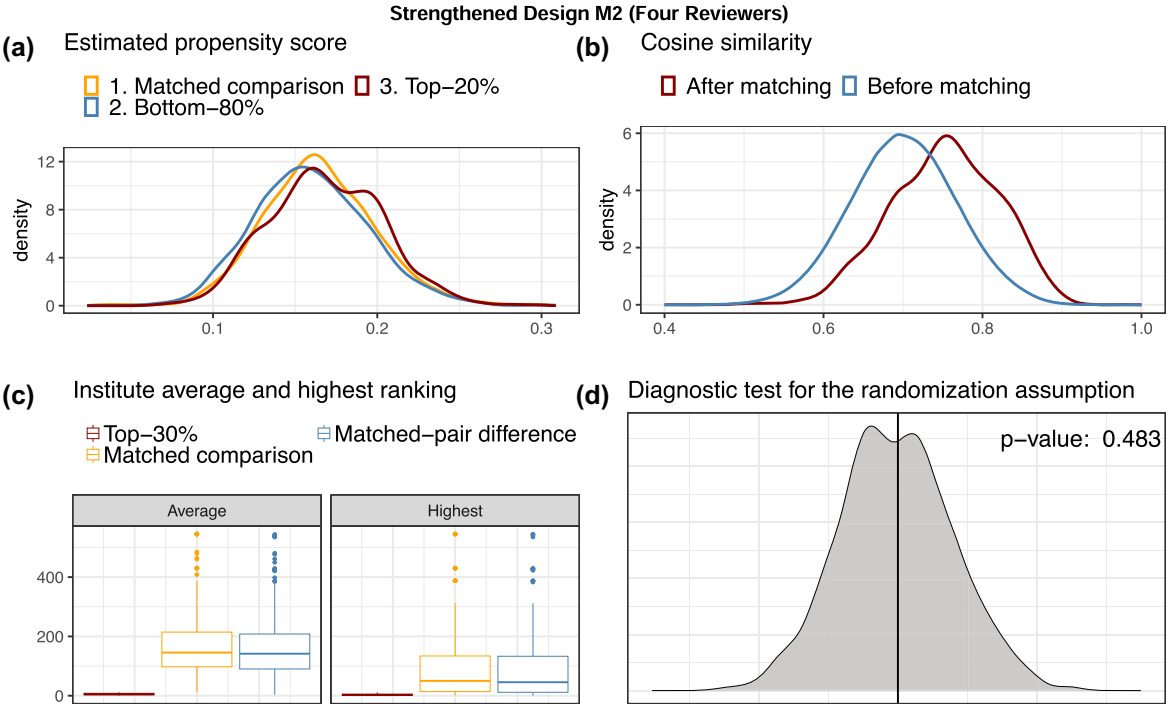

Figure A2: Diagnostics of design M2. Panel a: estimated propensity score top-30% articles, bottom-70% articles, and matched comparison articles; Panel b: distribution of between-article cosine similarity before (median = 0.70) and after matching (median = 0.76); Panel c: boxplots of authors' average (highest) institution rankings and matched-pair differences in authors' average (highest) institution rankings; Panel d: permutation distribution and the observed test statistic of the CPT.