# Publishing while female

## Are women held to higher standards? Evidence from peer review.*

Erin Hengel†

March 2020

(First version: September 2015)

Conditional on the quality of a paper, are women held to higher writing standards in academic peer review? Using readability scores to investigate, I find: (i) female-authored papers are 1–6 percent better written than equivalent papers by men; (ii) the gap widens *during* peer review; (iii) women improve their writing as they publish more papers (but men do not); (iv) men do not appear to compensate by raising quality along another dimension. Using a subjective expected utility framework, I show that tougher editorial standards are most obviously consistent with authors' observed choices. A conservative estimate derived from the model suggests higher writing standards may cause senior female economists to write at least 5–7 percent more clearly than they otherwise would.

*This paper is a revised version of the third chapter of my dissertation (University of Cambridge, September 2015). I am grateful to my supervisor Christopher Harris for (a) excellent guidance and (b) thinking this was a good idea. I am similarly indebted to Jeremy Edwards and my examination committee (Leonardo Felli and Hamish Low) for considerable input and advice. I also thank Miguel Almunia, Carolina Alves, Oriana Bandiera, Anne Boring, Cheryl Carleton, Gary Cook, Dominique Demougin, Harris Dellas, Carola Frege, Claudia Goldin, Olga Gorelkina, Jane Hunt, Ali Ismail, Adam Jaffe, Katya Kartashova, John Leahy, Brendan McCabe, Reshef Meir, Imran Rasul, Ludovic Renou, Kevin Schnepel, Joel Sobel, Heidi Williams, Jarrod Zhang, 20 (and counting!) anonymous referees and editors, audience members at the Econometric Society European Winter Meeting, the Eastern Economic Association Conference, the Royal Economic Society Annual Conference, the European Meeting of the Econometric Society, the American Economic Association Annual Meeting, the PEERE International Conference on Peer Review, the Review of Economics and Statistics Centenary Conference, the COSME Gender Economics Workshop, the Bank of England Conference on Gender and Career Progression, the NBER Summer Institute Innovation Workshop, the DIW Gender Economics Workshop, the German Economic Association Annual Meeting and seminar participants at the University of Liverpool, University of Cardiff, CEPII, Federal Reserve Board of Chicago, the NYU Center for Data Science, SOFI (Stockholm University), the University of Reading, the University of Sussex, London Business School, Diversifying Economics Network (University of Manchester), CRASSH (Cambridge University) and WERISE (Boston University). This paper could not have been written without substantial, careful research assistance by Michael Hengel (my dad), Eileen Hengel (my sister) and Lunna Ai (my actual research assistant). All errors, of course, are mine.

†University of Liverpool, Department of Economics; email: erin.hengel@liverpool.ac.uk.

## 1 Introduction

Ladies, our papers aren't published that often in "top-four" economics journals. In 2015, the average share of female authors per paper was 15 percent. Only eight percent were majority female-authored. Just four percent were written entirely by women. Between 2015–2017, the *Quarterly Journal of Economics* did not publish a single exclusively female-authored paper. In several recent years, *Econometrica* and the *Journal of Political Economy* have not either.

These statistics are uncomfortable, but their causes are myriad: lower publishing rates, career choices, motherhood and, possibly, bias. In lab and field experiments women are subject to tougher standards. Their qualifications and ability are underestimated (Foschi, 1996; Grunspan et al., 2016; Moss-Racusin et al., 2012; Reuben et al., 2014). Female-authored manuscripts are evaluated more critically (Goldberg, 1968; Krawczyk and Smyk, 2016; Paludi and Bauer, 1983). When collaborating with men, women are given less credit (Heilman and Haynes, 2005; Sarsons, 2019).

Although earlier studies haven't found much evidence of gender bias in peer review (see, *e.g.*, Blank, 1991; Borsuk et al., 2009; Gilbert et al., 1994; Lloyd, 1990), they tend to analyse a single indicator (acceptance rates) in a specific context (publication outcomes). In this paper, I ask a different question. Men's and women's manuscripts may be published at comparable *rates*, but are they scrutinised and evaluated by comparable *standards*? For example, if women are stereotypically assumed less capable at math, logic and reasoning than men and generally need more evidence to rate as equally competent, some referees may inspect their papers more closely, demand more revisions and have less patience deciphering their complicated, dense writing.

Complicated, dense writing is my focus. In the English language, clearly written prose is better prose, all things equal. Thoughtful word choice and simple sentence structure make text easier to understand, more interesting to read and expose inconsistencies long-winded writing often hides. Journal editors tend to agree. *Econometrica* asks authors to write "crisply but clearly" and to take "the extra effort involved in revising and reworking the manuscript until it will be clear to most if not all of our readers" (*Econometrica* submission guidelines, June 2016).[1]

To measure an article's writing clarity, I apply five highly tested "readability" formulas to 9,122 article abstracts published in the *American Economic Review* (*AER*), *Econometrica* (*ECA*), *Journal of Political Economy* (*JPE*) and *Quarterly Journal of Economics* (*QJE*).[2] To capture gender, I use each paper's proportion of female authors (see Section 2.1 for a justification), but also replicate most analyses using alternative ways to represent a paper's gender composition (see Appendix J).[3]

First, I find female-authored abstracts are 1–6 percent more readable than those by men. Women write better despite adjusting for other dimensions of quality—including citations, author prominence, seniority and individual fixed effects—accounting for English fluency and adding editor, journal, year and primary and tertiary *JEL* category dummies.

Second, the gender gap in readability is 2–3 times larger in the published version of a manuscript compared to its pre-submission version, even after conditioning on quality. Assuming authors do not make post-submission changes to their text unless requested by referees, these estimates suggest female-authored abstracts become 2–5 percent more readable because of peer review.

Third, the portion of the gap formed in peer review *reversed direction* in journals that blinded referees to authors' identities before the internet. Although standard errors are large and sample sizes small, this evidence tentatively suggests that blind review can mitigate the impact of gender under

[1] The *American Economic Review* rejected Robert Lucas's paper "Expectations and the Neutrality of Money" for insufficient readability; one referee wrote "If it has a clear result, it is hidden by the exposition" (Gans and Shepherd, 1994, p. 172). I additionally analysed 721 posts on Shit My Reviewers Say. A quarter deal with writing quality, document structure or word choice/tone.

[2] Abstract readability is highly correlated with the readability of other sections in a paper (see Appendix B.3 and Figure B.4).

[3] Where possible, I show results (a) on the subset of solo-authored papers; (b) on the subset of papers authored by a single gender; (c) using a binary variable equal to one if the most senior author was female; (d) using a binary variable equal to one if at least one author is female; and (e) using a binary variable equal to one if at least half of all authors are female.

certain circumstances. It also supports the hypothesis that editorial/refereeing bias is at least partially responsible for women's better writing.

Fourth, I do not find evidence that men compensate for their lower quality writing by raising quality along another dimension. Better writing by female economists could arguably compensate for some other advantage present in men's papers. But as long as men and women are equally capable researchers and similarly informed conditional on controls, the cost to both genders of implementing their respective publication strategies should be equal—otherwise, women could reduce the cost of producing a paper while holding acceptance rates constant by adopting a strategy marginally closer to men's (or visa versa). A rough test of this hypothesis using submit-accept times from *Econometrica* and the *Review of Economic Studies* (*REStud*) suggests this isn't the case. The cost to men of producing a paper appears to be much lower than the cost to women: female-authored papers spend ***three to six months longer*** in peer review compared to observably equivalent male-authored papers. The effect persists across a range of specifications and accounts for, among other things, citations, readability, author seniority, motherhood, childbirth and field.

Finally, it does not appear that women are rewarded for their better writing. Recent evidence at a set of four semi-overlapping journals suggests female-authored papers are *not* accepted at higher rates after conditioning on similar co-variates (Card et al., 2019).

Combined, these results suggest women spend too much time rewriting old papers and not enough time writing new papers, relative to men. The lack of a gender gap under blind review points to external factors beyond their control. But women's better writing could also be driven—or at least exacerbated—by internal factors such as higher risk-aversion (for a review, see Croson and Gneezy, 2009), lower confidence (see, *e.g.*, Coffman, 2014; Exley and Kessler, 2019), a tendency to update too much when faced with negative signals (Möbius et al., 2014), be more easily swayed by the opinions of others (Born et al., 2019) or exert more effort on low stakes tasks (Schlosser et al., 2019) and those do not yield obvious benefits (Babcock et al., 2017).

In order to determine wither external or internal factors primarily drive gender differences in readability, I model an author's decision-making process. Reviewers are assumed to accept papers only when their readabilities surpass author-specific minimum thresholds. Thresholds depend on non-readability aspects of papers—*e.g.*, more novel research is subject to lower thresholds—as well as referees' and editors' objectives, idiosyncratic preferences and biases and relative weight in determining outcomes. Authors then choose the actual readability of each of their papers in order to maximise a subjective expected utility that: (i) fixes their intrinsic preference for writing clearly over time; (ii) permits them to form misspecified beliefs about the relative importance of writing well; and (iii) allows the cost of writing to decline with experience.

The model suggests that if women improve their writing over time and are not commensurately rewarded with higher acceptance rates, then a gender readability gap between equivalent, experienced authors is primarily the result of holding women to higher writing standards. The intuition is simple. Assuming intrinsic preferences are fixed over time, authors improve their own writing only when they believe better writing leads to higher acceptance rates. Although poor information and oversensitivity may cause mistaken beliefs and mistaken beliefs can initially lead to suboptimal readability choices, authors correct such mistakes as they gain experience in peer review. Thus, a sufficiently experienced author writes more clearly than her inexperienced self only when writing clearly really does improve the probability her paper is accepted. If she also writes more clearly than an equivalent, experienced male author whose papers are accepted at rates no lower than hers, then higher standards—either in the form of biased referees or biased referee assignment—explain the difference (Theorem 1).

Theorem 1 establishes conditions sufficient to demonstrate double standards are present in academic peer review: (1) experienced women write better than equivalent men; (2) women improve their writing over time; (3) female-authored papers are accepted no more often than equivalent male-authored papers. Estimates from pooled subsamples at fixed publication counts suggest (1) and (2) hold. On average, women's writing gradually gets better but men's does not; between authors' first and third published articles, the readability gap increases by up to 12 percent. Although my data do not

identify Condition (3), female-authored papers are accepted *less* often than equivalent male-authored papers at a semi-overlapping set of journals (Card et al., 2019).

To interpret the relationship as causal, however, requires that Theorem 1's conditions hold for the same author. I therefore restrict the sample to authors with three or more top-four publications and match observably similar male and female economists based on characteristics—including citations and field—that predict the topic, novelty and quality of their research. Within-person readability comparisons determine if Condition (2) was satisfied for each author in a matched pair. Between-person comparisons after authors have gained experienced in peer review establishes whether Condition (1) was satisfied for the male or female member.

Conditions (1) and (2) were satisfied for the same author in 68 percent of matched pairs; in almost three-quarters of those, the member who satisfied them was female. Using a conservative estimate derived from the model, I estimate higher writing standards cause senior female economists to write at least 5–7 percent more clearly than they otherwise would.

I conclude by showing suggestive evidence that women navigate higher standards by altering their behaviour. Guided by the model, I tease out the direct effect of higher standards—readability changes made *in* peer review—from its "feedback" effect—readability changes made *before* peer review in anticipation of those higher standards—by comparing papers pre- and post-review as authors's gain experience in the process. In authors' earliest papers, the direct effect dominates. In fact, there is no significant gender difference between draft readabilities in men's and women's first top publications; it emerges entirely in peer review. In later papers, however, women write well upfront; the gap chiefly materialises *before* peer review.

These final results suggest women do not initially expect higher standards; instead, they learn about them over time and adapt their *ex ante* writing style accordingly. Consequently, papers by junior female economists likely experience the toughest review. Consistent with this hypothesis, I find a significantly smaller—albeit still positive—gender gap in peer review times for senior women.

This paper contributes to the literature in several ways. First, to the best of my knowledge, I am the first to suggest and document empirical evidence that women are held to higher standards in the peer review process (as opposed to its outcome). Higher standards have been recently corroborated using citations as a proxy for manuscript quality (Card et al., 2019; Grossbard et al., 2018; Hengel and Moon, 2019).[4] They also align with research on employee performance reviews, teaching evaluations and online comments: women receive more abusive feedback, less credit for intelligence and creativity and are expected to be more organised, prepared and clear (see, *e.g.*, Boring, 2017; Correll and Simard, 2016; Gardiner et al., 2016; Mengel et al., 2017; Wu, 2019).

Second, this paper proposes a novel explanation for academia's "Publishing Paradox", "Leaky Pipeline" and general promotion gap. Higher standards cause collateral damage to women's productivity: spending more time revising old research means there's less time for new research; fewer papers results in fewer promotions, possibly driving women into fairer fields.[5] They may also explain why so few women publish entirely female-authored papers, despite being the only work tenure committees give them full and fair credit for (Sarsons, 2019).

Third, my conclusions relate to a more general debate about gender differences in labour market outcomes.[6] Higher standards impose a quantity vs. quality trade-off that characterises female output

[4]Data from a field journal and 35 economics and finance journals also find female-authored manuscripts are subject to greater scrutiny and spend longer under review (Alexander et al., 2018; Hengel and Tol, 2018). A review-time gap was not, however, present in a set of journals that semi-overlap with those analysed here (Card et al., 2019).

[5]A similar idea was independently proposed in the philosophy literature (Bright, 2017). It was also informed by extensive research on editorial patterns (Card and DellaVigna, 2013; Card et al., 2019; Casnici et al., 2016; Clain and Leppel, 2018; Ellison, 2002), bias in editorial decisions (Abrevaya and Hamermesh, 2012; Bransch and Kvasnicka, 2017; Card and DellaVigna, 2017; Card et al., 2019) and female academics' lagging productivity and underrepresentation (Bayer and Rouse, 2016; Chari and Goldsmith-Pinkham, 2017; Ductor et al., 2018; Ginther and Kahn, 2004; Teele and Thelen, 2017).

[6]Traditional hypotheses focus on obvious discrimination (Goldin and Rouse, 2000), motherhood (Bertrand et al., 2010) and differences in behaviour (*e.g.*, Niederle and Vesterlund, 2010). Contemporary theories tend to stress inflexible working conditions (Goldin, 2014a; Goldin and Katz, 2016), preferences (for a review, see Blau and Kahn, 2017) and policy

in many professions—*e.g.*, doctors, real estate agents and airline pilots (for a discussion, see Hengel, 2017). Their downstream effects may contribute to several employment phenomena, including women's tendencies to concentrate in certain sectors and occupations (Blau and Kahn, 2017; Cortés and Pan, 2016), under-negotiate pay (Babcock and Laschever, 2003) and apply only to jobs they feel fully qualified for (Mohr, 2014). They may also reinforce work habits—*e.g.*, conscientiousness, tenacity and diligence—that correlate with quality and connote "femininity": for example, female physicians consult longer with patients (Roter and Hall, 2004), female politicians fundraise more intensely (Jenkins, 2007), female faculty commit fewer instances of academic misconduct (Fang et al., 2013) and female lawyers make fewer ethical violations (Hatamyar and Simmons, 2004).

Fourth, this paper joins an emerging body of economic research studying how the experience and anticipation of discrimination affects choices and behaviour. Earlier theoretical work focused on the impact discrimination has on investment in education and occupational choice (see, *e.g.*, Coate and Loury, 1993; Goldin, 2014b; Lundberg and Startz, 1983). More recent empirical research explores how stereotypes negatively impact performance (Bordalo et al., 2016; Carlana, 2019; Coffman, 2014; Glover et al., 2017; Lavy and Sand, 2015). My results suggest that rational responses to discrimination can distort productivity measurement (see also Parsons et al., 2011) and blur the line between biased treatment and voluntary choice.

Finally, this paper makes a related methodological contribution. Discrimination is generally identified from the actions (*e.g.*, Bertrand and Mullainathan, 2004; Neumark et al., 1996) and/or learning processes (*e.g.*, Altonji and Pierret, 2001; Fryer et al., 2013) of those who discriminate. But repeatedly observing authors' choices also exposes bias by editors and/or referees. In particular, multiple choices made under changing conditions reveals information about agents' intrinsic preferences and knowledge of underlying processes. Using this information, one can isolate group differences in the observed equilibrium from those that would have occurred in a non-discriminatory counterfactual one.[7] For example, assuming preferences are fixed over time, earlier choices provide an upper bound on the impact intrinsic preferences play in gender readability gaps; assuming authors update beliefs about the relationship between readability and acceptance rates means later choices are made with more accurate beliefs. Under modest assumptions, a similar strategy can be used in a variety of circumstances to credibly test for the existence of discrimination and identify—or at least bound—the effect it has on the long-term decision processes of those who experience it.

The remainder of the paper proceeds in the following order. Section 2 describes the data, the gender representation of articles published in top economics journals and readability scores. Analyses and results are presented in Section 3. Section 4 concludes.

## 2 Data

The data include every English-language article published with an abstract in *AER*, *ECA*, *JPE* and *QJE* between January 1950 and December 2015 (inclusive). The largest sample is from *Econometrica* which consistently published abstracts with its articles prior to 1950. *JPE* added them in the 1960s and *QJE* in 1980. *AER* came last in 1986. Unless otherwise mentioned, observations exclude the May issue of *AER*, *Papers & Proceedings* (*P&P*). Appendix B.2 displays data coverage by journal and decade.

For textual input, I use abstracts. Abstract readability is strongly positively correlated with the readability of other sections of a paper (see Figure B.4 and Hartley et al. (2003) and Plavén-Sigray et al. (2017)). Their structure is standardised in a manner optimal for computing readability scores. Most have also been converted to accurate machine readable text therefore curbing errors in transcription. See Appendix B.3 for further discussion.

design (Antecol et al., 2018).

[7] This idea is conceptually related to the infra-marginality problem. See Anwar and Fang (2006), Anwar and Fang (2015), and Knowles et al. (2001) for discussions in the context of racial discrimination.

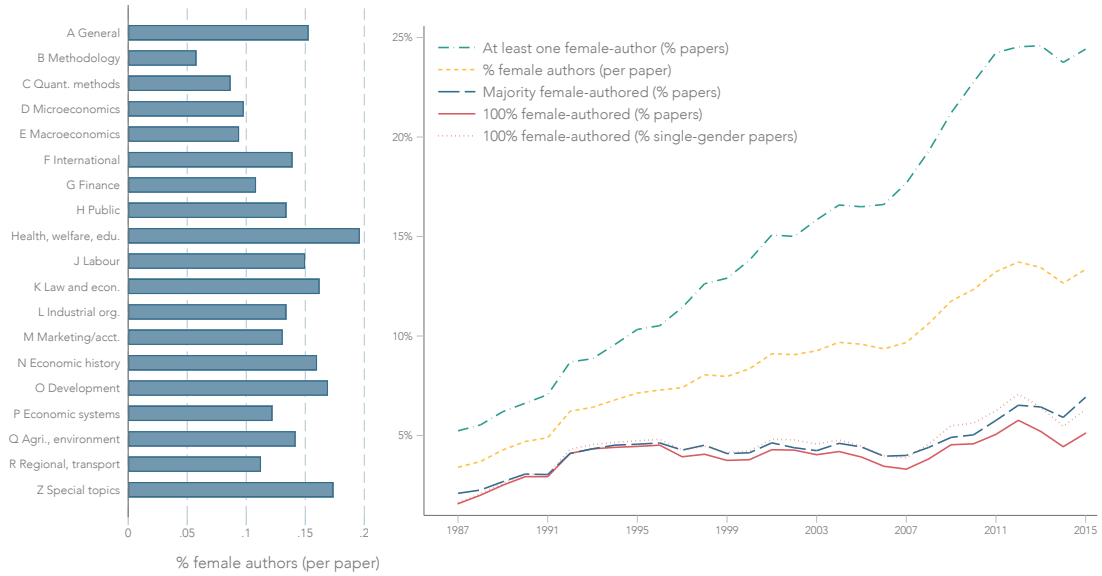


FIGURE 1: The representation of women in top economics journals

*Notes.* Graphs illustrate the representation of female authors in articles published in top-four economics journals. Figure on the left is the average share of female authors per paper broken down by primary *JEL* category (5,216 articles); figure on the right displays the evolution of papers' gender composition over time (6,176 articles; estimates are five-year moving averages).

For the analysis in Section 3.2, I collected draft abstracts from NBER Technical and Working Paper Series. To match published articles with their NBER drafts, I used citation data from RePEc and searched NBER's database directly for unmatched papers authored by NBER family members. 1,986 published articles were eventually matched to 1,988 NBER working papers—approximately one-fifth of the data.[8] Descriptive statistics are shown in Section 3.2.2.

The analysis in Section 3.4 compiles submit-accept times at *Econometrica* (1970–2015) and *REStud* (1976–2015), a fifth highly respected economics journal; *AER*, *JPE* and *QJE* do not make disaggregated data on their revision process publicly available. I obtained the data from journals' online archives or extracted it from digitised articles using the open source command utility `pdftotext`. Section 3.4 displays and discusses basic summary statistics.

Other variables used in the analysis include editor fixed effects, dynamic institution fixed effects, primary and tertiary *JEL* fixed effects, controls for author prominence—total lifetime number of top-five (top-four plus *REStud*) articles for the most prolific co-author—and seniority—total number of top-five articles at time of publication for the most-prolific co-author—English fluency dummies, citation counts (asinh), and controls for motherhood and childbirth (Section 3.4, only). See Appendix C for further information on how each was calculated.

### *2.1 Gender*

Authors were assigned a gender using GenderChecker.com's database of male and female names. Authors with unisex first names, first names not in the database or those identified only by initial(s) were assigned a gender either by me, a research assistant or at least three separate Mechanical Turk workers based on a visual inspection of photos on faculty websites, Wikipedia articles, *etc.* or personal pronouns used in text written about the individual. In situations where the author could not be found but several people with the same first and last name were and all shared the same gender, the author was also assigned that gender. For the remaining cases, I emailed or telephoned colleagues and institutions associated with the author.

[8] The mapping is not one-for-one because a small number of working papers were eventually published as multiple articles or combined into one.

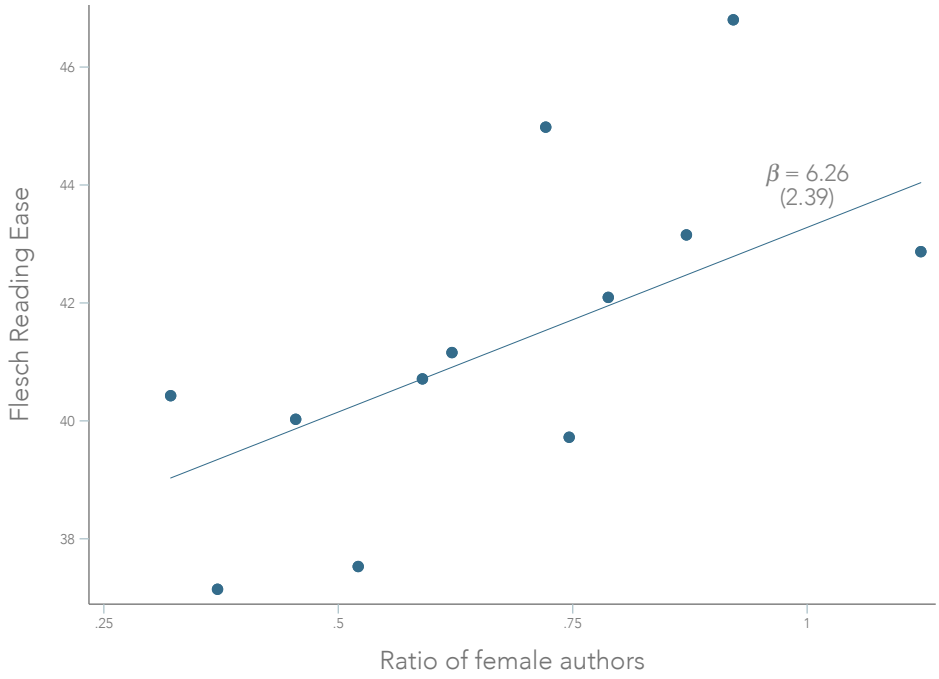


FIGURE 2: Relationship between readability and ratio of female authors

*Notes.* Binned scatter plot of abstract readability against the ratio of female authors on papers with at least one female author (1,166 articles; sample excludes three articles with six or more authors, only one of which was female). Estimates control for whether the paper was solo- or co-authored (resulting in a small proportion of articles having a female ratio above one).

Determining the "gender" of a paper is not nearly as straightforward. For solo-authored manuscripts—of which there are 4,016 in the sample—gender corresponds to the sex of the author. Unfortunately, top economics journals have collectively published just 266 by women. Only a slightly larger number were written entirely—*or even mostly*—by women (Figure 1).[9]

Instead, I represent an article's gender using its proportion of female authors. First, this allows me to take advantage of the information contained in the much larger sample of papers authored by at least one woman (1,172). Second, a gender readability gap—if it exists—is presumably a function of (i) the probability a passage of text was written and/or revised by a female co-author; and (ii) referees' beliefs about female authors' contributions to the writing and/or revision of a co-authored paper. Prior research suggests co-authors—regardless of seniority—share responsibility for writing and (especially) revising collaborative work (see, *e.g.*, Hart, 2000; Kumar and Ratnavelu, 2016). Thus, the intersection of (i) and (ii) is likely positively related to the ratio of female authors on a paper.[10]

Figure 2 corroborates this hypothesis. It plots papers' abstract readability against their ratio of female authors. The slope of the regression line is positive, relatively large (6.26 points on the Flesch Reading Ease scale) and highly statistically significant.

For robustness, however, I repeat most analyses (a) on the sample of solo-authored papers, only; (b) comparing papers with a senior female co-author to entirely male-authored papers; (c) on the subset of papers authored by a single gender; (d) using a binary variable equal to one if at least one author is female; and (e) using a binary variable equal to one if at least half of all authors are female. Standard errors from (a) and (c) tend to be larger; those from (b), (d) and (e) usually smaller. In general, however, results do not meaningfully change (Appendix J).

[9] 312 papers in the sample were authored entirely by women. Women made up more than 50 percent of all authors in another 47. In 35 observations, a woman was the lead author—*i.e.*, the first author was female in a paper with authors listed non-alphabetically or in which contributions were explicitly noted.

[10] In Appendix H, I find evidence suggesting the relationship is increasing and convex.

TABLE 1: Textual characteristics per sentence, by gender

| | Men | Women | Difference |
|---|---|---|---|
| No. characters | 134.83 | 131.39 | −3.44** |
| | (0.43) | (1.31) | (1.38) |
| No. words | 24.19 | 23.39 | −0.80*** |
| | (0.08) | (0.24) | (0.25) |
| No. syllables | 40.69 | 39.16 | −1.53*** |
| | (0.13) | (0.40) | (0.42) |
| No. polysyllabic words | 4.70 | 4.40 | −0.30*** |
| | (0.02) | (0.06) | (0.06) |
| No. difficult words | 9.39 | 9.03 | −0.36*** |
| | (0.03) | (0.10) | (0.11) |

*Notes.* Sample 8,800 articles. Figures are means of textual characteristics (per sentence) by sex. Male means are of exclusively male-authored papers (7,948 articles); female means are for majority female-authored papers (female ratio at or above 50 percent) (852 articles). Last column subtracts male means from female means. Standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

### *2.2 Readability scores*

To measure writing clarity, I use the five most common, widely tested and reliable readability formulas for adult-level material: Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG (Simple Measure of Gobbledegook) and Dale-Chall (see Figure B.1, Appendix B). Appendix B.1 reviews the literature on readability score validity. Appendix B.2 breaks abstract readability down by publication year and primary *JEL* classification. Appendix B.3 discusses measurement error and the impact it potentially has on results and conclusions.

The Flesch Reading Ease formula ranks passages of text in ascending order—*i.e.*, more readable passages earn higher scores. The other four formulas, however, generate grade levels estimating the minimum years of schooling necessary to confidently understand an evaluated text—and so more readable passages earn lower scores. In order to simplify interpretation, I multiple the four grade-level scores by negative one. Thus, higher scores universally correspond to clearer writing throughout this paper.

To calculate the scores, I wrote the Python module `Textatistic`. Its code and documentation are available on GitHub; a brief description is provided in Appendix B.4. For added robustness, I also re-calculate scores and replicate most results using the `R readability` package (Appendix J).

## 3 Analyses and results

### *3.1 Gender differences in readability*

Table 1 compares textual characteristics between male- and female-authored papers. It suggests women write shorter, simpler sentences: they contain fewer characters, fewer syllables, fewer words and fewer "hard" words. Differences are highly statistically significant.

Table 2 presents coefficients from 40 separate ordinary least squares (OLS) regressions of readability scores on the ratio of female authors. Column (1) includes journal and editor fixed effect.[11] Columns (2) and (3) add journal, year and journal-year interaction dummies. Column (4) introduces controls for paper $j$'s number of co-authors ($N_j$) and the dynamic institution effects described in Appendix C. Column (5) adds a dummy variable capturing English fluency; it also controls for article quality (citations (asinh)), co-author prominence (max. $T_5^j$) and seniority at the time of publication

[11] The coefficients on the journal dummies in (2) are presented in Appendix D. Compared to *AER*, all five scores agree that *Econometrica* is harder to read; four out of five scores suggest *JPE* is, too, while *QJE* is easier.

TABLE 2: Gender differences in readability, article-level analysis

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.90* | 0.87* | 0.83* | 0.89* | 1.14** | 0.80 | 0.79 | 0.99 |
| | (0.48) | (0.48) | (0.50) | (0.50) | (0.49) | (0.55) | (0.55) | (0.70) |
| Flesch-Kincaid | 0.18 | 0.17 | 0.17 | 0.18 | 0.21* | 0.24* | 0.26** | 0.26* |
| | (0.11) | (0.11) | (0.11) | (0.11) | (0.12) | (0.13) | (0.12) | (0.14) |
| Gunning Fog | 0.31** | 0.31** | 0.31** | 0.33*** | 0.36*** | 0.40** | 0.38*** | 0.37** |
| | (0.12) | (0.12) | (0.12) | (0.12) | (0.13) | (0.15) | (0.14) | (0.16) |
| SMOG | 0.20** | 0.20** | 0.20** | 0.21** | 0.24*** | 0.25** | 0.23** | 0.24* |
| | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.11) | (0.10) | (0.12) |
| Dale-Chall | 0.10** | 0.10** | 0.10** | 0.10** | 0.12** | 0.13** | 0.12** | 0.15** |
| | (0.04) | (0.04) | (0.05) | (0.05) | (0.05) | (0.06) | (0.06) | (0.06) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* Sample 9,117 articles in (1)–(5); 5,211 articles in (6) and (7); 5,774 articles—including 563 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Figures represent coefficients from 40 separate OLS regressions of readability scores on the ratio of female authors. Coefficients in (6) are estimated using the controls in (5) but on the sample from (7). Quality controls denoted by ✓[1] include citation count (asinh), max. $T_5$ fixed effects (author prominence) and max. $t_5$ (author seniority). Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

(max. $t_5$). Columns (6)–(8) are estimated on the sample of articles published after 1990. (7) and (8) include fixed effects for primary and tertiary *JEL* categories, respectively.[12]

Results in Table 2 suggest that abstracts written by women score about one point higher on the Flesch Reading Ease scale; according to the four grade-level measures, they take about 2–3 fewer months of schooling to understand. Percentage-wise, women write about 1–2 percent better than men.

Appendix D.2 explores field in more detail; Appendix H analyses readability at the author-level. Conditional on other explanatory variables, I find little evidence that field drives results in Table 2. After accounting for author-specific heterogeneity, the gender gap in readability is 2–6 percent.

### *3.2 The causal impact of peer review*

**3.2.1 Identification.** Comparing abstracts pre- and post-review make it possible to isolate gender differences in readability pre-existing peer review from those incurred during it—and therefore identify the immediate effect of gender inside peer review. In this section, I present two different methodologies that estimate this effect.

The first strategy simply regresses each paper's change in score on its gender composition. To understand it, note that the readability of a published paper depends on its earlier draft readability as

[12] Due to small sample sizes, column (8) includes 561 articles from *AER P&P*, coded as a separate journal. Papers published in *AER P&P* are selected and edited by the American Economic Association's president-elect with the help of a Program Committee (see www.aeaweb.org for more details). *P&P* does not publish abstracts in its print version; only select years (2003 and 2011–2015) and papers were available online when I collected the data (first in early 2015 and then updated in early 2016). Excluding these articles does not impact results or conclusions: coefficients are similar to those in column (8), but standard errors are somewhat higher.

well as factors that affect writing clarity any time *after* it was initially drafted:

$$R_{jP} = R_{jW} + \beta_{0P} + \beta_{1P}\,\text{female ratio}_j + \boldsymbol{\theta}_P \mathbf{X}_{jP} + \mu_{jP} + \varepsilon_{jP}, \tag{1}$$

where $R_{jP}$ and $R_{jW}$ are readability scores for working ($W$) and published ($P$) versions of paper $j$, respectively. $\beta_{1P}$ is the coefficient of interest and reflects the particular impact female ratio$_j$ has in peer review. $\mathbf{X}_{jP}$ and $\mu_{jP}$ are $P$-specific observable and unobservable components, respectively. $\varepsilon_{jP}$ is $P$'s error term.

Correlation between $R_{jW}$ and female ratio$_j$ biases OLS estimates of $\beta_{1P}$. Equation (2) eliminates the distortion by subtracting $R_{jW}$ from both sides of Equation (1):

$$R_{jP} - R_{jW} = \beta_{0P} + \beta_{1P}\,\text{female ratio}_j + \boldsymbol{\theta}_P \mathbf{X}_{jP} + \mu_{jP} + \varepsilon_{jP}. \tag{2}$$

Assuming zero partial correlation between female ratio$_j$ and $\mu_{jP}$, OLS generates an unbiased estimate of $\beta_{1P}$.[13]

An alternative strategy based on Ashenfelter and Krueger (1994) separately estimates gender differences in the draft and final versions of papers using generalised least squares (GLS). The contemporaneous effect of peer review is identified post-estimation by subtracting coefficients. To implement this set-up, I need to additionally define the relationship between readability scores and the gender composition of a paper *before* peer review:

$$R_{jW} = \beta_{0W} + \beta_{1W}\,\text{female ratio}_j + \boldsymbol{\theta}_W \mathbf{X}_{jW} + \mu_{jW} + \varepsilon_{jW}, \tag{3}$$

where $\beta_{1W}$ reflects female ratio$_j$'s impact on readability prior to peer review; $\mathbf{X}_{jW}$ and $\mu_{jW}$ are version-invariant observable and unobservable components, respectively; $\varepsilon_{jW}$ is version $W$'s error term. Equation (4) then defines a general structure for potential correlation between $\mu_{jW}$ and observable variables in both Equation (3) and Equation (1):

$$\mu_{jW} = \gamma + \eta\,\text{female ratio}_j + \boldsymbol{\delta}_W \mathbf{X}_{jW} + \boldsymbol{\delta}_P \mathbf{X}_{jP} + \omega_j, \tag{4}$$

where $\omega_j$ is uncorrelated with female ratio$_j$, $\mathbf{X}_{jW}$ and $\mathbf{X}_{jP}$. Substituting Equation (4) into Equation (3) generates the following reduced form representation of $R_{jW}$:

$$R_{jW} = \widetilde{\beta}_{0W} + \widetilde{\beta}_{1W}\,\text{female ratio}_j + \widetilde{\boldsymbol{\theta}}_W \mathbf{X}_{jW} + \boldsymbol{\delta}_P \mathbf{X}_{jP} + \widetilde{\varepsilon}_{jW}, \tag{5}$$

where $\widetilde{\beta}_{0W} = \beta_{0W} + \gamma$, $\widetilde{\beta}_{1W} = \beta_{1W} + \eta$, $\widetilde{\boldsymbol{\theta}}_W = \boldsymbol{\theta}_W + \boldsymbol{\delta}_W$ and $\widetilde{\varepsilon}_{jW} = \varepsilon_{jW} + \omega_j$. $R_{jP}$'s reduced form is similarly found by substituting Equation (5) into Equation (1):

$$R_{jP} = (\widetilde{\beta}_{0W} + \beta_{0P}) + (\widetilde{\beta}_{1W} + \beta_{1P})\,\text{female ratio}_j + \widetilde{\boldsymbol{\theta}}_W \mathbf{X}_{jW} + \widetilde{\boldsymbol{\theta}}_P \mathbf{X}_{jP} + \mu_{jP} + \widetilde{\varepsilon}_{jP}, \tag{6}$$

where $\widetilde{\boldsymbol{\theta}}_P = \boldsymbol{\theta}_P + \boldsymbol{\delta}_P$ and $\widetilde{\varepsilon}_{jP} = \widetilde{\varepsilon}_{jW} + \varepsilon_{jP}$. Equation (5) and Equation (6) are explicitly estimated via feasible GLS (FGLS). $\beta_{1P}$ is identified post-estimation by subtracting reduced form coefficients.

Both OLS estimation of Equation (2) and FGLS estimation of Equation (5) and Equation (6) require zero partial correlation between $\mu_{jP}$ and female ratio$_j$ to obtain a valid $\beta_{1P}$. Roughly restated, non-peer review factors must be either independent of its timing or unrelated to gender.[14] Section 3.2.4 evaluates this assumption.

[13] Note that Equation (2) implicitly controls for all factors—*e.g.*, research field—that impact draft readability but are not otherwise affected by peer review.

[14] This phrasing is slightly inaccurate but convenient for exposition. Zero correlation between female ratio$_j$ and $\mu_{jP}$ does not preclude biased estimates of $\beta_{1P}$ when $\mu_{jP}$ is correlated with other explanatory variables that are, in turn, correlated with female ratio$_j$ by some factor independent of $\mu_{jP}$. Unbiasedness instead requires zero *partial* correlation between $\mu_{jP}$ and female ratio$_j$.

TABLE 3: Textual characteristics, published papers vs. drafts

| | Men | | | Women | | | |
|---|---|---|---|---|---|---|---|
| | Working paper | Published article | Difference | Working paper | Published article | Difference | Diff.-in diff. |
| No. sentences | 6.45 | 5.07 | −1.374*** | 6.77 | 5.06 | −1.711*** | −0.337** |
| | (0.06) | (0.04) | (0.057) | (0.15) | (0.08) | (0.139) | (0.149) |
| No. characters | 860.09 | 647.26 | −212.826*** | 907.36 | 635.97 | −271.385*** | −58.558*** |
| | (7.56) | (4.93) | (7.523) | (18.53) | (10.31) | (18.439) | (19.597) |
| No. words | 155.34 | 115.31 | −40.034*** | 164.45 | 113.63 | −50.813*** | −10.779*** |
| | (1.39) | (0.90) | (1.393) | (3.42) | (1.91) | (3.428) | (3.630) |
| No. syllables | 256.36 | 192.68 | −63.682*** | 269.02 | 187.78 | −81.242*** | −17.560*** |
| | (2.25) | (1.48) | (2.242) | (5.54) | (3.08) | (5.504) | (5.843) |
| No. polysyllabic words | 28.31 | 21.76 | −6.557*** | 28.93 | 20.63 | −8.308*** | −1.751*** |
| | (0.29) | (0.19) | (0.257) | (0.71) | (0.41) | (0.627) | (0.668) |
| No. difficult words | 58.37 | 44.48 | −13.897*** | 60.32 | 42.37 | −17.949*** | −4.052*** |
| | (0.54) | (0.35) | (0.507) | (1.30) | (0.74) | (1.204) | (1.315) |
| Flesch Reading Ease | 41.40 | 41.11 | −0.298 | 42.51 | 43.08 | 0.564 | 0.862* |
| | (0.28) | (0.19) | (0.193) | (0.66) | (0.43) | (0.452) | (0.500) |
| Flesch-Kincaid | −13.65 | −13.40 | 0.249*** | −13.53 | −13.00 | 0.531*** | 0.282** |
| | (0.06) | (0.05) | (0.052) | (0.15) | (0.11) | (0.122) | (0.134) |
| Gunning Fog | −17.32 | −17.06 | 0.252*** | −17.13 | −16.58 | 0.547*** | 0.295** |
| | (0.07) | (0.05) | (0.057) | (0.18) | (0.13) | (0.140) | (0.149) |
| SMOG | −15.16 | −15.02 | 0.138*** | −15.02 | −14.70 | 0.327*** | 0.189* |
| | (0.05) | (0.04) | (0.037) | (0.13) | (0.09) | (0.095) | (0.097) |
| Dale-Chall | −10.85 | −10.93 | −0.082*** | −10.71 | −10.70 | 0.003 | 0.085** |
| | (0.02) | (0.02) | (0.016) | (0.06) | (0.04) | (0.037) | (0.042) |

*Notes*. Sample 1,566 published articles authored entirely by men (1,567 NBER working papers); 272 published articles authored by at least 50 percent women (273 NBER working papers). Figures are means of textual characteristics by sex for NBER working papers and published articles. Penultimate columns in each panel subtract working paper figures from published article figures for men (first panel) and women (second panel); difference-in-differences (female less male) shown in the final column. Standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

**3.2.2 Summary statistics.** Draft abstracts were collected from NBER Technical and Working Paper Series (see Section 2). NBER series were used as the exclusive data source for two reasons. First, approximately one-fifth of articles in my data were originally part of an NBER series, making it the largest single source of draft papers. Second, authors release their manuscripts as NBER working papers at about the same time they submit them to a journal (see Ellison, 2002; Goldberg, 2015, and Figure E.1).

Table 3 compares textual characteristics between a paper's draft and final versions. It suggests abstract text is altered during peer review. According to the first panel, draft abstracts are longer—more characters, words and sentences—and denser—more syllables, polysyllabic words and difficult words. The biggest changes are made to female-authored papers: figures in column six are 20–30 percent higher (in absolute value) than those in column three. The second panel of Table 3 suggests peer review improves readability, although results are less clear for male-authored papers.

**3.2.3 Results.** Table 4's first column displays $\beta_{1P}$ from OLS estimation of Equation (1). Conditional on draft readability, published female-authored papers are more readable than published male-authored papers. Moreover, published article readability positively correlates with draft readability: coefficients on $R_{jW}$ (shown in Appendix E.1) are positive and significant—but only about 0.8. The less than unit value suggests $\mu_{jP}$ exerts downward pressure on $R_{jW}$'s coefficient, thereby artificially inflating first column figures.

Table 4's remaining columns show results from the two strategies presented in Section 3.2.1 to deal with this bias. The FGLS strategy estimates the coefficient on female ratio$_j$ separately among the

TABLE 4: The impact of peer review on the gender readability gap

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.31** | 2.21** | 3.15*** | 0.94 | 0.94 |
| | (0.58) | (0.98) | (1.21) | (0.59) | (0.60) |
| Flesch-Kincaid | 0.52*** | 0.32 | 0.76*** | 0.44** | 0.44** |
| | (0.17) | (0.22) | (0.28) | (0.19) | (0.19) |
| Gunning Fog | 0.51*** | 0.44* | 0.85*** | 0.41** | 0.41** |
| | (0.19) | (0.24) | (0.30) | (0.19) | (0.20) |
| SMOG | 0.30** | 0.32** | 0.56*** | 0.24** | 0.24* |
| | (0.13) | (0.15) | (0.19) | (0.12) | (0.12) |
| Dale-Chall | 0.18*** | 0.32*** | 0.44*** | 0.13** | 0.13** |
| | (0.05) | (0.10) | (0.11) | (0.05) | (0.05) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,709 NBER working papers; 1,707 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles. Column one displays coefficients on female ratio ($\beta_{1P}$) from estimating Equation (1) directly via OLS (see Appendix E.1 for coefficients on $R_{jW}$); standard errors clustered by editor in parentheses. Columns two and three display $\widetilde{\beta}_{1W}$ and $\widetilde{\beta}_{1W} + \beta_{1P}$ from FGLS estimation of Equation (5) and Equation (6), respectively; standard errors clusterd by year and robust to cross-model correlation in parentheses. Their difference ($\beta_{1P}$) is shown in column four. Column five displays $\beta_{1P}$ from OLS estimation of Equation (2); standard errors clustered by year in parentheses. Quality controls denoted by ✓[2] include citation count (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

sample of working papers (column two) and published articles (column three). The impact of gender on the readability gap formed in peer review is the difference between them (column four). Combined, results in columns 2–4 suggest women's better writing is indeed caused by peer review: the gender readability gap is 2–3 times larger in papers' published versions than it was in their pre-print versions; percentage-wise, immediate peer review accounts for about 30–60 percent of the final gap.

The OLS strategy regresses each paper's change in score on its gender composition. As discussed in Section 3.2.1, this specification has the added benefit of completely removing the impact of confounding factors—*e.g.*, research field—that are constant between versions. Coefficients on female ratio$_j$ are shown in Table 4's fifth column. Their magnitudes and standard errors almost perfectly mirror FGLS estimates.

As shown in Appendix E.1, the relationship between citations and the *change* in readability between draft and final versions of a paper is either negative or zero.[15] Although we do not observe how many citations papers would have received had they not gone through peer review, these results tentatively suggest that the revisions women are asked to make in peer review may not improve the quality of their papers.

*Double-blind review.* To estimate the impact double-blind review had on the gender readability gap, I add the dummy variable blind$_j$ and its interaction with female ratio$_j$ to Equation (2):

$$\begin{aligned} R_{jP} - R_{jW} = \beta_{0P} + \beta_{1P}\,\text{female ratio}_j + \beta_{2P}\,\text{blind}_j + \beta_{3P}\,\text{female ratio}_j \times \text{blind}_j \\ + \boldsymbol{\theta}_P \mathbf{X}_{jP} + \mu_{jP} + \varepsilon_{jP}, \end{aligned} \tag{7}$$

[15] Citations and abstract readability do, however, (generally) positively correlate (see Appendix B).

TABLE 5: The impact of blind peer review

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Blind ($\beta_{1P} + \beta_{3P}$) | −1.52 | −0.57 | −0.55 | −0.37 | −0.12 |
| | (2.93) | (0.67) | (0.79) | (0.57) | (0.17) |
| Non-blind ($\beta_{1P}$) | 0.94 | 0.44** | 0.41** | 0.24* | 0.13** |
| | (0.60) | (0.19) | (0.20) | (0.12) | (0.05) |
| Difference ($\beta_{3P}$) | −2.46 | −1.00 | −0.96 | −0.61 | −0.25 |
| | (3.01) | (0.72) | (0.83) | (0.59) | (0.17) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 1,988 NBER working papers; 1,986 published articles. Columns display marginal effects on female ratio for papers undergoing non-blind ($\beta_{1P}$) and blind ($\beta_{1P} + \beta_{3P}$) review from OLS estimation of Equation (7). Standard errors clustered by year in parentheses. Quality controls denoted by ✓[2] include citation count (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

where blind$_j$ is equal to 1 if article $j$ was subjected to an official policy of double-blind review before Google incorporated in 1998.[16]

Table 5's first two rows display marginal effects of female ratio under non-blind ($\beta_{1P}$) and blind ($\beta_{1P}+\beta_{3P}$) review from OLS estimation of Equation (7). They suggest a smaller—possibly negative—gap under blinded peer review. Marginal effects in single-blind reviewed papers are identical to figures in Table 4.

Table 5's final row reports differences between effects ($\beta_{3P}$). Their consistent positive direction provides some (weak) evidence that masking authors' identities reduces peer review's impact on the gender readability gap. The effect, however, did not survive the internet. In Appendix E.3, I analyse the policy's post-internet impact. Gender differences are positive regardless of a journal's official review policy, suggesting that double-blind review is effective only as long as authors are not identifiable by other means.

**3.2.4 Robustness.** Timing independence is the principle assumption required to causally link the readability gap to the peer review process. Post-submission, manuscripts probably only change because of peer review—either because referees request specific changes or authors believe, possibly mistakenly (see Section 3.3), that they will be requested in a future revision. Thus, timing independence is arguably only violated during the narrow timeframe after a manuscript is released as an NBER Working Paper but before it is submitted to a top-four journal. As Appendix E.4 illustrates, only a small proportion of papers are exposed to this window: most manuscripts—and especially most female-authored

[16] Three notes on this definition. First, double-blind review was likely less effective after the internet was adopted (for anecdotal evidence, see, *e.g.*, Goldberg, 2014). I therefore only evaluate the impact of blind review pre-internet. See Appendix E.3 for an analysis of the policy's post-internet impact. Second, from 1 May 1987 to 31 May 1989, half of all papers submitted to *AER* were evaluated by single-blind review; the remaining half were subjected to double-blind review (for details on the trial, see Blank, 1991). Referees correctly identified at least one author in 45.6 percent of double-blind reviewed papers—indicating that only about a quarter of the manuscripts were truly blind reviewed. I therefore classify every paper published during the trial as having undergone single-bind review. (Note that excluding these observations has almost no impact on results.) Third, as discussed in Blank (1991), a final publication date may substantially lag the actual review date. Nevertheless, results are unchanged when including only *AER* articles published post May 1989 and all *QJE* articles published before June 2005 were evaluated under double-blind review (*Econometrica* and *JPE* have never blinded referees to authors' identities). Thus, misclassification errors of this kind are unlikely to substantially bias estimates presented in Table 5.

manuscripts—are submitted to peer review at the same time or *before* they are released as NBER Working Papers.

Another concern is that gender differences in how authors conform to abstract word limits may bias results in Table 4. To investigate this possibility, I exclude the 642 observations—about 40 percent of the sample—with NBER abstracts longer than the official word limit of the journals in which they were eventually published. Results are presented in Appendix E.5. Coefficient magnitudes are similar to those in Table 4; standard errors are somewhat larger.

Finally, in an effort to maximise sample sizes, estimates in the first three columns of Table 4 omit field controls. Including them slightly increases standard errors; they otherwise make little difference (see Appendix E.2). Estimates in the final column implicitly account for field already (see Footnote 13).

### *3.3 Distinguishing between causal mechanisms*

**3.3.1 Theoretical framework.** The previous section suggests that female-authored abstracts become 2–5 percent more readable *because of* peer review. Two causal mechanisms could explain this link: either (i) women voluntarily write better papers—*e.g.*, because they're more sensitive to referee criticism or their papers are more innovative and therefore harder to understand—or (ii) better written papers are women's response to higher standards imposed by referees and/or editors.

To distinguish between (i) and (ii), I develop a simple model of an author's decision making process. It follows an author—denoted by $i$—who publishes several articles in prestigious academic journals over the course of his career. Each article is roughly equivalent in terms of topic, novelty and quality, but may vary on readability.

At stage 0, author $i$ drafts his $t$th paper and submits it for peer review. Upon receipt, the journal's editorial office assigns the manuscript to a group of referees. The (finite) set of all potential review groups is represented by $\Sigma$; $\mu_i$ is the set of strictly positive probability measures on $\Sigma$. $\Sigma$ and $\mu_i$ are known to $i$.

$\widetilde{r}^s_{0i}$ is the readability threshold below which review group $s \in \Sigma$ rejects any paper by author $i$. I assume it depends on other qualities of $i$'s papers—*e.g.*, methodological rigour, data, originality, policy relevance, *etc.* If $i$'s papers are strong in these characteristics, $\widetilde{r}^s_{0i}$ will be low (and vice versa). $\widetilde{r}^s_{0i}$ may also reflect reviewers' objectives, idiosyncratic preferences and relative weight in determining outcomes. For example, an editor who does not care about readability and is willing to override the opinion of referees will implement a lower $\widetilde{r}^s_{0i}$ (all else equal).

Conditional on non-readability characteristics such as topic, novelty and overall quality, $s$ rejects $i$'s paper at stage 0 if

$$r_{0it} < \widetilde{r}^s_{0i},$$

where $r_{0it}$ is manuscript $t$'s draft readability. $i$ is otherwise granted a "revise and resubmit" (R&R), yet could still be rejected at stage 1 if the readability of his revised manuscript, $R_{it} = r_{0it} + r_{1it}$, does not meet a second threshold,

$$R_{it} < \widetilde{R}^s_i,$$

where $\widetilde{R}^s_i = \widetilde{r}^s_{0i} + \widetilde{r}^s_{1i}$. All rejections and acceptances are final. $\widetilde{R}^s_i \neq \widetilde{r}^s_{0i}$ to account for different standards at different stages of peer review. $r_{0it}$, $r_{1it}$, $\widetilde{r}^s_{0i}$ and $\widetilde{r}^s_{1i}$ are non-negative; the latter two are independent.

To aid the revision process, $s$ writes a referee report from which $i$ forms expectations about $\widetilde{R}^s_i$ by assigning subjective probabilities $\pi^s_{1it}(R)$ to all $R$. Unfortunately, the concept of readability is complex, some referees write insufficiently detailed reports and inattentive or hypersensitive authors misconstrue even perfectly clear advice. This renders $i$'s interpretation of the report imprecise and potentially biases his subsequent expectations about $\widetilde{R}^s_i$.

Conditional on $r_{0it}$, I assume referee reports by $s$ for $i$ are the same for all $t$ and that each is distinctive enough for $i$ to distinguish $s$ in $\Sigma$.[17] Consequently, author $i$'s stage 1 choice of $R_{it}$ maximises his (immediate) subjective expected utility given $s$,

$$\Pi^s_{1it}(R_{it})u_i + \phi_{i|r_{0it}}(r_{1it}) - c_{it|r_{0it}}(r_{1it}). \tag{8}$$

$\Pi^s_{1it}(R_{it})$ is the cumulative sum of $\pi^s_{1it}(R)$ for all $R \leq R_{it}$; $u_i$ is the utility of having a paper accepted in a prestigious journal.[18]

$\phi_{i|r_{0it}}(r_{1it}) = \phi_i(R_{it}) - \phi_i(r_{0it})$ and $c_{it|r_{0it}}(r_{1it}) = c_{it}(R_{it}) - c_{it}(r_{0it})$ are the satisfaction and cost, respectively, $i$ derives from making changes $r_{1it}$ at time $t$ given the paper's initial readability $r_{0it}$. $\phi_i$ is increasing and concave in its arguments, $c_{it}$ increasing and convex—marginally higher $R_{it}$ generates proportionally less satisfaction but needs more effort when the paper is already well written. Thanks to learning, the cost of writing clearly may fall over time ($\mathrm{d}c_{it}/\mathrm{d}t \leq 0$); however, it converges uniformly to the limiting function $c_i$—*i.e.*, learning is "smooth" over $t$.

Authors' decisions at stage 0 are myopic; $i$'s choice of $r_{0it}$ maximises his initial subjective expected utility for the current paper,

$$\int_\Sigma \Pi^s_{0it}(r_{0it})v^s_{1it}\,\mathrm{d}\mu_i + \phi_i(r_{0it}) - c_{it}(r_{0it}), \tag{9}$$

where $\Pi^s_{0it}(r_{0it})$ is the cumulative sum for all $r \leq r_{0it}$ of author $i$'s subjective probabilities $\pi^s_{0it}(r)$ about $\widetilde{r}^s_{0i}$; $v^s_{1it}$ is Equation (8) evaluated at the optimal $r_{1it}$.

Authors update subjective probabilities (i) using relevant information from their own experience in peer review; and (ii) by observing others' readability choices and publication outcomes. When evidence from (i) contradicts evidence from (ii), (i) takes precedence. These assumptions imply, at a minimum, that $i$ updates $\Pi^s_{0it}$ and $\Pi^s_{1it}$ based on conclusive evidence derived from the choices and outcomes of equivalent peers (Definition 1) and knowledge acquired from his own experience in peer review.

**Definition 1.** *Equivalent authors write papers that are identical with respect to topic, novelty and quality.*

Equation (8) and Equation (9) incorporate a variety of factors that potentially affect authors' readability choices—editorial standards conditional on other qualities in the paper ($\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$); ambition ($u_i$); the cost of drafting and revising manuscripts ($c_{it}$); an otherwise unexplained intrinsic satisfaction from writing readable papers ($\phi_i$). Poor information, overconfidence and sensitivity to criticism are not explicitly included, on the assumption that people do not *want* to be poorly informed, overconfident or excessively sensitive. These factors nevertheless enter Equation (8) and Equation (9)—and hence influence choices—via the subjective expectations authors form about $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$.

A single $R_{it}$ cannot, therefore, establish if and to what extent $i$'s choices are motivated by (a) preferences and costs specific to him ($u_i$, $\phi_i$, $c_{it}$), (b) conditional editorial standards and/or referee assignment outside his control ($\widetilde{r}^s_{0i}$, $\widetilde{R}^s_i$, $\mu_i$) or (c) miscellaneous confounding factors mopped up by $\Pi^s_{0it}$ and $\Pi^s_{1it}$. Since $i$'s preferences and limiting cost function ($c_i$) are time independent, however, observing an increase in his choice of readability at two separate $t$ distinguishes (a) from the combined impact of (b) and (c):[19] $i$ may be more sensitive to criticism and he might prefer writing more clearly; nevertheless, he improves readability today relative to yesterday only when he believes it boosts his chances of publishing. Moreover, because (c) does not reflect activities or states the author enjoys, its

[17] Should $s$ review a future paper by $i$, $i$ would recognise it as the same group that reviewed his earlier paper. This does not imply that the report reveals individual referees' identities.

[18] Authors probably care about getting their papers accepted and they may care about writing well, but their marginal utility from the intersection of the two events—*i.e.*, higher utility from writing well *only* because the paper is published in a top-four journal (as opposed to a top field journal or second-tier general interest journal)—is assumed to be negligible.

[19] The analysis in Section 3.2 similarly establishes that (b) and/or (c) are significant factors driving the choice of $R_{it}$. It cannot, however, distinguish between them (although the double-blind review analysis points to (b)).

impact on choices declines with experience—*i.e.*, authors may miscalculate referee expectations and misconstrue their reports, but with experience they correct those mistakes.

I capture this idea in Theorem 1, where $\mathbf{1}^s_{0i}(r)$ and $\mathbf{1}^s_{1i}(R)$ are indicator functions equal to 1 if $r \geq \widetilde{r}^s_{0i}$ and $R \geq \widetilde{R}^s_i$, respectively, and $\Sigma_{A_{it}}$ is the collection of $s \in \Sigma$ for which $\mathbf{1}^s_{0i}(r_{0it})\mathbf{1}^s_{1i}(R_{it}) = 1$. Theorem 1 is proved in Appendix A.

**Theorem 1.** *Consider two equivalent authors, $i$ and $k$, that satisfy the following three conditions.*

*Condition 1.* $(r_{0kt}, R_{kt}) \leq (r_{0it}, R_{it})$ *for all* $s \in \Sigma_{A_{it}}$ *and* $t > t'$ *and there exists* $K' > 0$ *such that for at least one* $s \in \Sigma_{A_{it}}$ *and no* $t > t'$, $||(r_{0it}, R_{it}) - (r_{0kt}, R_{kt})|| < K'$.

*Condition 2.* *For at least one* $t'' < t'$, $(r_{0it''}, R_{it''}) < (r_{0it'}, R_{it'})$ *and there exists* $K'' > 0$ *such that for no* $t > t'$, $||(r_{0it}, R_{it}) - (r_{0it''}, R_{it''})|| < K''$.

*Condition 3.* $\int_\Sigma \mathbf{1}^s_{0i}(r_{0it})\mathbf{1}^s_{1i}(R_{it})\,\mathrm{d}\mu_i \leq \int_\Sigma \mathbf{1}^s_{0k}(r_{0kt})\mathbf{1}^s_{1k}(R_{kt})\,\mathrm{d}\mu_k$ *for all* $t > t'$.

*Then, almost surely, referee assignment is biased in favour of $k$,*

$$\int_\Sigma \mathbf{1}^s_{0i}(r_{0kt})\mathbf{1}^s_{1i}(R_{kt})\,\mathrm{d}\mu_i < \int_\Sigma \mathbf{1}^s_{0i}(r_{0kt})\mathbf{1}^s_{1i}(R_{kt})\,\mathrm{d}\mu_k,$$

*or referee scrutiny is biased against $i$,*

$$\int_\Sigma \mathbf{1}^s_{0i}(r_{0kt})\mathbf{1}^s_{1i}(R_{kt})\,\mathrm{d}\mu_i < \int_\Sigma \mathbf{1}^s_{0k}(r_{0kt})\mathbf{1}^s_{1k}(R_{kt})\,\mathrm{d}\mu_i,$$

*or both.*

Theorem 1's three conditions are sufficient to verify discrimination is present in academic publishing. That is, when female authors' unconditional probability of acceptance is no higher than men's (Condition 3), their current papers are more readable than their past papers (Condition 2) and also *persistently* more readable than men's papers (Condition 1) then either editors assign women "tougher" referees—*i.e.*, those with higher $\widetilde{r}^s_{0i}$ and/or $\widetilde{R}^s_i$—or they apply higher standards to women's writing—*i.e.*, $\widetilde{r}^s_{0k} < \widetilde{r}^s_{0i}$ and/or $\widetilde{R}^s_k < \widetilde{R}^s_i$ for at least one $s \in \Sigma$.

*Measuring higher standards.* Theorem 1 principally relies on two identifying assumptions: (i) $i$ and $k$ satisfy Definition 1 at time $t'$; (ii) $t'$ is sufficiently large—*i.e.*, any errors in $i$'s beliefs about $\widetilde{r}_{0i}$ and $\widetilde{R}_i$ are on a path converging to zero. Corollary 1 assumes a more specific belief structure at $t'$ in order to (conservatively) measure discrimination's impact on readability choices.

Let $e^s_{0it}$ and $e^s_{1it}$ be $i$'s time $t$ error in beliefs about $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$, respectively, and define $\overline{s}$ as the review group in $\Sigma_{A_{it}}$ for which $i$ believes $\widetilde{r}^s_{0i}$ is highest. If $i$ and $k$ are equivalent at $t > t'$ and $e^s_{nit} = e^s_{nkt}$ at stages $n = 0, 1$, then Corollary 1 shows that $R_{it} - R_{kt}$ is *smaller* in magnitude than the true value of stage 1 discrimination by $s$ $(\widetilde{R}^s_i - \widetilde{R}^s_k)$ or stage 0 discrimination by $\overline{s}$ $(\widetilde{r}^{\overline{s}}_{0i} - \widetilde{r}^{\overline{s}}_{0k})$.

**Corollary 1.** *Fix $s$ and $t > t'$ and suppose (i) $i$ and $k$ are equivalent authors such that $i$ satisfies Conditions 1–3 (Theorem 1) relative to $k$; (ii) $e^s_{nit} = e^s_{nkt}$ for stages $n = 0, 1$; and (iii) $\Sigma_{A_{it}} \subset \Sigma_{A_{kt}}$. Then*

$$\underline{D}_{ik} \equiv R_{it} - R_{kt} \leq D_{ik}, \tag{10}$$

*where*

$$D_{ik} \equiv \begin{cases} \widetilde{R}^s_i - \widetilde{R}^s_k & \text{if } r_{0it} < R_{it} \\ \widetilde{r}^{\overline{s}}_{0i} - \widetilde{r}^{\overline{s}}_{0k} & \text{otherwise} \end{cases}.$$

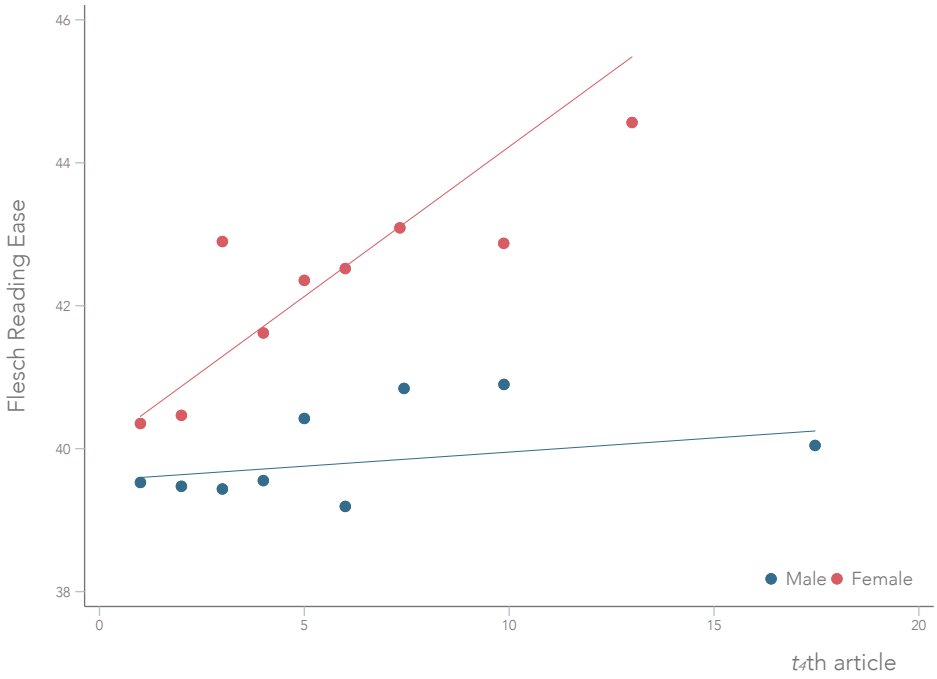


FIGURE 3: Readability of authors' $t_4$th top-four publication

*Notes.* Binned scatter plot of abstract readability for authors' first, second, …, $t_4$th, … top-four publication. Sample 1,306 female observations (767 disctinct female authors) and 14,759 male observations (6,109 distinct male authors).

Corollary 1 conservatively measures the impact of discrimination on $i$'s readability. It also exposes the toxic influence of a single biased $s$: $i$'s time $t$ readability choice depends on discrimination at stage 1 by the group of referees that actually reviewed his paper ($s$) as well as discrimination at stage 0 by another review group that (probably) didn't ($\overline{s}$).

Corollary 1 adds two stronger conditions to Theorem 1. First, $i$ and $k$ must be comparably experienced by time $t$.[20] Second, if $s' \in \Sigma_{A_{it}}$ then $s' \in \Sigma_{A_{kt}}$. This second condition might not be satisfied if, *e.g.*, $i$'s utility of acceptance exceeds that of $k$'s so he works harder to appease the demands of a particularly tough group of reviewers. Nevertheless, $i$'s unconditional acceptance rate is not higher than $k$'s (Condition 3), so there must also exist some other $s''$ that applies higher standards to $i$'s work than it does to $k$'s (thus, $s'' \in \Sigma_{A_{kt}}$ but $s'' \notin \Sigma_{A_{it}}$). However, Equation (10) may not fully counteract the first effect relative to the second (see the proof in Appendix A). Equation (11) does. It therefore conservatively estimates $D_{ik}$ when $\Sigma_{A_{it}} \not\subset \Sigma_{A_{kt}}$:

$$\underline{D}_{ik} \equiv R_{it} - \max\{R_{it''}, R_{kt}\} \leq D_{ik}, \tag{11}$$

where $t'' < t$ is defined in Condition 2 of Theorem 1.[21]

### 3.3.2 Empirical results.

*Descriptive evidence.* In this section, I show suggestive, non-causal evidence that female authors satisfy Theorem 1's three conditions relative to male authors.

Consider first Condition 3: female-authored papers are accepted no more often than male-authored papers. The articles I evaluate in this paper have already been published, precluding gender analysis of acceptance rates. Nevertheless, the topic has been extensively studied elsewhere. A recent study of four journals that semi-overlap with my own suggests exclusively male- and female-authored manuscripts

[20] Corollary 1 actually applies under the weaker $e^s_{nit} \leq e^s_{nkt}$, $n = 0, 1$ (see its proof in Appendix A).
[21] Equation (11) does come with a cost: its conservative bias is much larger than the one generated by Equation (10).

receive a revise and resubmit decision 8 and 6 percent of the time, respectively (Card et al., 2019). Blank (1991) found that 12.7 and 10.6 percent of male- and female-authored papers were accepted at the *AER*, respectively. A study of *JAMA*'s editorial process indicated that 44.8 percent of referees accept male-authored papers as is or if suitably revised; 29.6 percent summarily reject them. Corresponding figures for female-authored papers were 38.3 and 33.3 percent, respectively (Gilbert et al., 1994). Studies from other disciplines find female-authored papers subjected to single-blind peer review are accepted less often than would be expected by chance (Handley et al., 2015; McGillivray and De Ranieri, 2018). There appear to be no gender differences in acceptance rates to NBER's Summer Institute (Chari and Goldsmith-Pinkham, 2017). Desk rejection rates may actually be higher for female-authored papers submitted to the field journal *Energy Economics* (Tol, 2018). Ceci et al. (2014) provide a more comprehensive research review on the subject. Their conclusion: "When it comes to actual manuscripts submitted to actual journals, the evidence for gender fairness is unequivocal: there are no sex differences in acceptance rates." (Ceci et al., 2014, p. 111).

My data better identify Conditions 1 and 2. As their careers advance, women write more clearly: their average readability scores are 1–5 percent higher than the readability of their first papers, their latest papers 1–7 percent higher; for a man, however, his average and last papers are more poorly written than his first (Appendix F.1, Table F.1). Figure 3 suggests a similar story. It plots an author's Flesch Reading Ease score against $t_4$, where $t_4 = 1$ for his first top-four publication, $t_4 = 2$ for his second, *etc.* As $t_4$ increases, men's and women's readability diverges.

Table 6 tests the significance of that divergence, conditioning on confounders. It presents marginal effects on co-authoring with women for female authors ($\beta_1$) from estimating Equation (12) on subsamples of authors where $t_4 = 1$, $t_4 = 2$, *etc.*:

$$R_{j_{it_4}} = \beta_0 + \beta_1 \text{ female ratio}_j + \text{female ratio}_j \times \text{male}_i + \boldsymbol{\theta}\,\mathbf{X}_j + \varepsilon_{it_4}, \tag{12}$$

where $R_{j_{it_4}}$ is the readability score for article $j$, author $i$'s $t_4$th top-four publication. Gender enters twice—the binary variable $\text{male}_i$ and $\text{female ratio}_j$—to account for $i$'s sex and the sex of his co-authors. $\mathbf{X}_j$ is a vector of observable controls and $\varepsilon_{it}$ is the error term.[22]

All figures in Table 6 agree—women write better—but the magnitude and significance of that difference increases as $t_4$ increases. Between $t_4 = 1$ and $t_4 = 2$, the gap marginally widens but is not significant; after that, it triples (at least); the increase is significant ($p < 0.05$) for all five scores (Appendix F.2, Table F.2). At higher publication counts, figures are less precisely estimated and somewhat smaller than in column 3—but still noticeably larger than estimates in columns 1 and 2.[23]

*Estimation strategy.* Evidence in the previous section suggests women satisfy Theorem 1's three conditions relative to men, on average. Yet the set of women to satisfy one condition is conceivably orthogonal to sets that satisfy others; for Theorem 1 to apply, they must overlap.

To address this concern, I restrict the sample to authors with three or more top-four publications and match observably similar male and female economists based on characteristics—including citations and field—that predict the topic, novelty and quality of their research. In addition to explicitly accounting for author equivalence—the primary conditional independence assumption behind Theorem 1—matched pair comparisons: (i) identify the gender most likely to satisfy all conditions simultaneously; and (ii) generate (conservative) estimates of the effect higher standards have on authors' readability (Corollary 1).

Holding acceptance rates constant, Theorem 1 rules out confounding factors—*e.g.*, sensitivity to criticism and individual preferences—by comparing readability between equivalent authors experienced in peer review (Condition 1) and within authors before and after gaining that experience (Condition 2). I consider authors "experienced" by $t_4 = 3$. Authors with one or two top-four publications are probably

[22] Data used in Figure 3 and Table 6 were disaggregated to the author-level by duplicating each article $N_j$ times. To account for duplicate articles, regressions in Table 6 are weighted by $1/N_j$.

[23] Only 40 female authors have 4–5 publications in the data; 28 have six or more.

TABLE 6: Gender gap in readability at increasing $t_4$

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch Reading Ease | 0.45 | 1.49* | 4.99*** | 3.01 | 2.31 | 1.88** |
| | (0.68) | (0.84) | (1.13) | (1.96) | (2.14) | (0.73) |
| Flesch-Kincaid | 0.08 | 0.14 | 0.96*** | 0.66* | 0.47 | 0.23 |
| | (0.15) | (0.19) | (0.22) | (0.40) | (0.37) | (0.15) |
| Gunning Fog | 0.22 | 0.32 | 1.30*** | 0.95** | 0.66 | 0.45** |
| | (0.17) | (0.24) | (0.27) | (0.44) | (0.43) | (0.19) |
| SMOG | 0.14 | 0.24 | 0.86*** | 0.70** | 0.47 | 0.35*** |
| | (0.12) | (0.17) | (0.19) | (0.35) | (0.30) | (0.13) |
| Dale-Chall | 0.07 | 0.11 | 0.38*** | 0.29* | 0.39* | 0.19*** |
| | (0.06) | (0.08) | (0.12) | (0.16) | (0.23) | (0.07) |
| No. observations | 6,875 | 2,826 | 1,675 | 1,906 | 2,773 | 12,006 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* $\beta_1$ from FGLS estimation of Equation (12). First column restricts sample to authors' first top-four publication ($t_4 = 1$), second column to their second ($t_4 = 2$), *etc.* Regressions weighted by $1/N_j$ (see Footnote 22). Standard errors (in parentheses) adjusted for two-way clustering (editor and author) and cross-model correlation. Final column estimates from an unweighted population-averaged regression; error correlations specified by an auto-regressive process of order one and standard errors (in parentheses) adjusted for one-way clustering on author. Quality controls denoted by ✓[1] include citation count (asinh), max. $T_5$ fixed effects (author prominence) and max. $t_5$ (author seniority); ✓[3] includes citation count (asinh) and max. $t_5$, only. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

tenured and well-established in their fields. By publication three, all frequently referee (and some edit) prestigious economics journals. I assume this accumulated experience means equivalent authors are equally accurate about $\widetilde{r}_{0i3}$ and $\widetilde{R}_{i3}$; remaining errors are no longer gender specific: $e^s_{ni3} = e^s_{nk3}$, $n = 0, 1$ (Corollary 1).[24]

To account for equivalence, I match every female author with three or more top-four publications (121) to her closest male counterpart (1,554). Matches were made using a Mahalanobis procedure with the following co-variates: (1) maximum citation count over $t_4$; (2) institutional rank at $t_4 = 1$; (3) fraction of papers published per decade; (4) fraction of papers published by each journal; and (5) number of articles per primary *JEL* category.[25] Co-variate balance pre- and post-match are shown in Appendix F.3. Appendix F.4 lists each matched pair.

Under ideal circumstances, $R_{i3} - R_{i1}$ measures the impact experience has on readability, conditional on gender; $R_{i3} - R_{k3}$ measures gender's impact conditional on experience. Because of co-authoring, however, article gender is neither fixed over $t_4$ conditional on $i$, nor is its difference constant between $i$ and $k$, conditional on $t_4$.[26] To account for this, I create a counterfactual $\widehat{R}_{it}$ that captures $i$'s $t_4$th paper readability had it only been co-authored with members of $i$'s same sex. It is reconstructed at female ratio equal to 1 for women and 0 for men using errors and coefficients from OLS estimation of

[24] Nevertheless, $e^s_{nit} - e^s_{nkt}$ converges to 0 as $t$ tends to infinity. Thus, $\underline{D}_{ik}$ will consistently predict the *direction* of $D_{ik}$ for a sufficiently large $t$ even when errors remain gender-specific.

[25] Two notes on co-variate choice. First, I eschew mean, median and minimum citation counts in favour of the maximum on the assumption that an author's "quality" is principally a function of his best paper. Second, most people are at top ranked institutions by $t_4 = 3$; by matching on $t_4 = 1$ institution, I try to pair authors with similar career paths. The robustness of results to these and other co-variate choices are discussed in a following section.

[26] One way authors can meet higher standards is by co-authoring with better writers. Thus, co-authorship by itself is not a confounding factor.

TABLE 7: $\underline{D}_{ik}$, Equation (10)

| | Discrimination against women ($\underline{D}_{ik} > 0$) | | | Discrimination against men ($\underline{D}_{ik} < 0$) | | | Mean, all observations | |
|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | $N$ | Mean | S.D. | $N$ | (1) | (2) |
| Flesch Reading Ease | 13.46 | 11.06 | 60 | −8.37 | 8.70 | 20 | 5.03*** | 3.99*** |
| | | | | | | | (1.14) | (1.23) |
| Flesch Kincaid | 2.87 | 2.18 | 65 | −2.66 | 2.46 | 19 | 1.15*** | 1.00*** |
| | | | | | | | (0.26) | (0.27) |
| Gunning Fog | 3.50 | 2.76 | 62 | −2.61 | 2.89 | 21 | 1.37*** | 1.18*** |
| | | | | | | | (0.31) | (0.33) |
| SMOG | 2.73 | 1.96 | 54 | −1.40 | 1.94 | 24 | 0.87*** | 0.71*** |
| | | | | | | | (0.22) | (0.23) |
| Dale-Chall | 1.38 | 0.93 | 62 | −0.94 | 0.69 | 22 | 0.53*** | 0.43*** |
| | | | | | | | (0.11) | (0.12) |

*Notes.* Sample 121 matched pairs (109 and 121 distinct men and women, respectively). First and second panels display conditional means, standard deviations and observation counts of $\underline{D}_{ik}$ (Equation (10)) from subpopulations of matched pairs in which the woman or man, respectively, satisfies Conditions 1 and 2. Third panel displays mean $\underline{D}_{ik}$ over all observations. To account for the 30–40 percent of pairs for which Theorem 1 is inconclusive, (1) sets $\underline{D}_{ik} = 0$, while (2) sets $\underline{D}_{ik} = \widehat{R}_{i3} - \widehat{R}_{k3}$ if $\widehat{R}_{i3} < \widehat{R}_{k3}$ ($i$ female, $k$ male) and zero, otherwise. Male scores are subtracted from female scores; $\underline{D}_{ik}$ is positive in panel one and negative in panel two. $\underline{D}_{ik}$ weighted by frequency observations are used in a match; degrees-of-freedom corrected standard errors in parentheses (panel three, only). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Equation (13) in the gender and time appropriate subsample of authors:[27]

$$\widehat{R}_{it} = \alpha_{tg_i} + \beta_{tg_i} \text{ female ratio}_{it} + \widehat{\varepsilon}_{it}, \tag{13}$$

where $g_i = m, f$ if $i$ is male or female, respectively and $\widehat{\varepsilon}_{it}$ is the estimated error term. Regression output from Equation (13) is shown in Appendix F.5. To adjust for the degrees of freedom lost when generating $\widehat{R}_{it}$, standard errors in subsequent calculations are inflated by 1.05.

As long as $\widehat{\varepsilon}_{it}$ does not partially correlate with a paper's ratio of female authors conditional on $t$ and $g_i$, then $\widehat{R}_{it}$ provides an unbiased prediction of $R_{it}$.[28] The validity and robustness of this and other assumptions relevant to Theorem 1 and Corollary 1 are discussed in a following section.

*Results.* Table 7 tests if Conditions 1 and 2 are both satisfied within each matched pair. Its first and second panels display the mean and standard deviation of $\underline{D}_{ik}$ (Equation (10)) and observation counts from the set of matched pairs in which one member satisfies both conditions. In the first panel, the female member does, suggesting discrimination against women. In the second, it's the male member (indicating discrimination against men). Male scores are subtracted from female scores, so $\underline{D}_{ik}$ is, by definition, positive in panel one and negative in panel two.

Evidence of discrimination was present in 68 percent of matched pairs. In almost three-quarters of those, the member discriminated against was female. Moreover, $\underline{D}_{ik}$ is (on average) 1.5 times as large (in absolute value) when discrimination is against women.

Figure 4 displays $\underline{D}_{ik}$'s distribution across the five scores. In the absence of systemic discrimination against women (or men), $\underline{D}_{ik}$ would symmetrically distribute around zero. It does not. When men are discriminated against, $\underline{D}_{ik}$ clusters closer to zero. When women are discriminated against, $\underline{D}_{ik}$ is more spread out. Furthermore, instances of obvious discrimination are predominately against women: $\underline{D}_{ik}$ is five times more likely to be one standard deviation above zero than below it.

[27] More specifically, I separately estimate Equation (13) in the following four subsamples: (i) female authors at $t_4 = 1$; (ii) male authors at $t_4 = 1$; (iii) female authors at $t_4 = 3$; (iv) male authors at $t_4 = 3$. I then generate $\widehat{R}_{it}$ using the appropriate coefficients and errors for each author: (i) $\widehat{R}_{i1} = \alpha_{1f} + \beta_{1f} + \widehat{\varepsilon}_{i1}$ for a female $i$ at $t_4 = 1$; (ii) $\widehat{R}_{i1} = \alpha_{1m} + \widehat{\varepsilon}_{i1}$ for a male $i$ at $t_4 = 1$; *etc.*

[28] That is female ratio$_{it}$ cannot act as a collider on the causal chain from higher standards to greater readability, conditional on the gender and experience of an author. Appendix F.8 replicates relevant analyses without adjusting for female ratio$_{it}$.

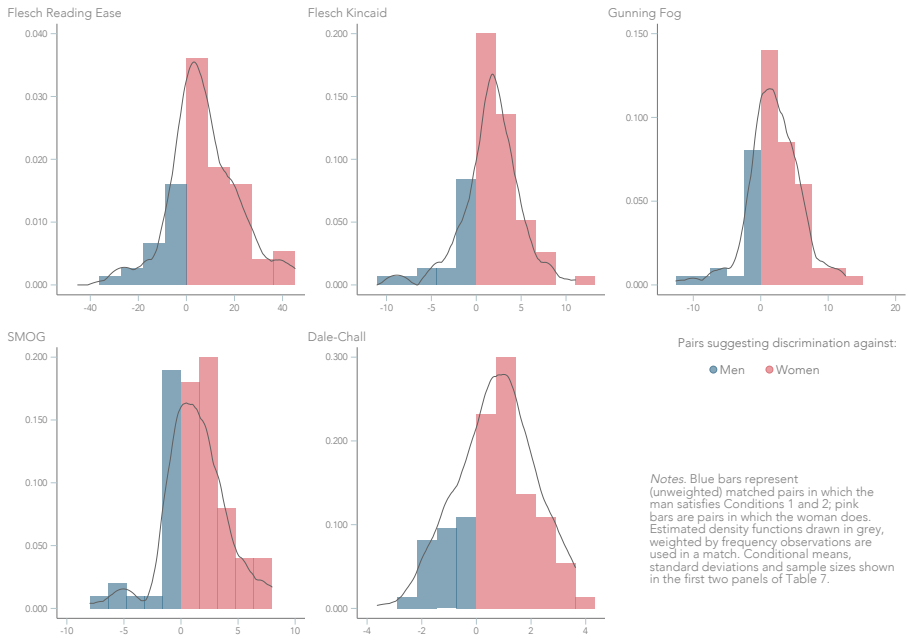


FIGURE 4: Distributions of $\underline{D}_{ik}$, Equation (10)

Table 7's final panel averages $\underline{D}_{ik}$ over all observations. To account for the 30–40 percent of pairs for which neither member satisfies both Conditions 1 and 2, (1) sets $\underline{D}_{ik} = 0$, whereas (2) sets $\underline{D}_{ik} = \widehat{R}_{i3} - \widehat{R}_{k3}$ if $\widehat{R}_{i3} < \widehat{R}_{k3}$ ($i$ female, $k$ male) and zero, otherwise.[29] Mean $\underline{D}_{ik}$ is positive and significant in both columns for all five scores. On average, first column figures suggest that higher standards cause senior female economists to write (at least) seven percent more clearly than they otherwise would.

Appendix F.6 replicates Table 7 using $\underline{D}_{ik}$ from Equation (11). Estimates are, by definition, smaller—they suggest higher standards cause female economists to write (at least) five percent more clearly than they otherwise would—but conclusions are unchanged.

*Robustness.* Conclusions drawn from Table 7 and Figure 4 are principally predicated on two assumptions: (1) $i$ and $k$ are equivalent at $t_4 = 3$; and (2) $t_4 = 3$ is sufficiently large. If either is violated, discrimination against women cannot be inferred from an overrepresentation of matched pairs with $\underline{D}_{ik} > 0$.

Assumption (1) depends on match accuracy. Post-match co-variates are well balanced (see Appendix F.3). They remain well balanced—and similar to the matched population—when restricted to pairs satisfying $\underline{D}_{ik} \neq 0$. To facilitate further scrutiny, Appendix F.4 lists the names of economists in each pair. Matches using alternative variables (*e.g.*, minimum citation counts, mean institutional rank or fraction of articles per primary *JEL* category) and specifications (*e.g.*, propensity score matching) generate similar figures and conclusions.[30]

Assumption (2) demands a "sufficiently large" $t_4$. For diagnosing discrimination, "sufficiently large" means $t' < 3$ and the difference in $i$ and $k$'s error in beliefs at $t_4 = 3$ is smaller than $D_{ik}$. Fifty percent of women with three or more top publications satisfy Conditions 1 *and* 2 when compared to equivalent men. Among them, $\underline{D}_{ik}$ is far from zero: these women write, on average, 21 percent more clearly than equivalent men with identical experience. I believe it is unlikely that half of all female economists with

[29] That is, if the experienced man writes more readably than the experienced woman, then the effect is always attributed to discrimination against men; if the experienced woman writes more readably than the experienced man, however, the effect is attributed to discrimination against women only if Condition 2 is likewise satisfied.

[30] See Hengel (2017, pp. 30–33) for propensity score matches from a probit model performed with replacement and using a wider array of co-variates; results from alternative matching algorithms are available on request.

three top publications—plus many more second-tier publications and substantial experience refereeing and editing themselves—make mistakes of this magnitude.

To generate the counterfactual $\widehat{R}_{it}$ (Equation (13)), I assume unobserved co-author characteristics do not partially correlate with female ratio$_{it}$, conditional on $i$'s gender and experience. To test the robustness of this assumption, Table J.27 (Appendix J.5) replicates Table 6 on exclusively, majority and senior female-authored papers. I have also repeated the analyses shown in Table 7 and Figure 4 without adjusting for female ratio$_{it}$ (Appendix F.8) and on subsets of matched pairs in which the woman's $t_4 = 1$ and $t_4 = 3$ papers are solo- or exclusively female-authored (16), majority female-authored (20) or at least 50 percent female-authored (76). Although sample sizes for the latter three analyses are small, they also find $\underline{D}_{ik} \neq 0$ in about 70–75 percent of matched pairs; most of those (70 percent) indicate discrimination against the female member; the impact across all five scores also averages about 7 percent.

Moreover, experience appears to be the only $t_4$-varying factor driving within $i$ changes in readability. Table 6 and additional analyses in a 2016 version of this paper (Hengel, 2016, pp. 23–24) show an identical pattern despite controlling for a large array of potential confounders. In a 2017 version, I reconstructed $\widehat{R}_{it}$ using several $t_4$-varying factors (number of co-authors, institutional rank, institutional rank of the highest ranked co-author, $t_4$ for the most experienced co-author, publication year and dummies for each journal) (Hengel, 2017, pp. 30, 61); Appendix F.7 adds *JEL* classification codes to Equation (13). In Table J.27 (Appendix J.5), I restrict Table 6's analysis to solo-authored papers or those co-authored by members of the same sex. In all instances, women's readability is consistently shown to increase with $t_4$; when comparable results are estimated, they are similar to those presented in Table 7 and Figure 4.

Finally, causal interpretation technically requires that three additional criteria are also met. Assuming discrimination against $i$: (i) $i$'s acceptance rate is no more than $k$'s; (ii) $i$'s draft readability is at least as high as $k$'s; and (iii) $i$'s draft readability at $t_4 = 3$ is at least as high as his draft readability at $t_4 = 1$. As already discussed, (i) rules out the possibility that $i$ is appropriately rewarded (relative to $k$) for writing more clearly. (ii) and (iii) eliminate situations in which women write more clearly during peer review in order to compensate for poorer writing—and consequently higher desk rejection rates—before peer review.

Unfortunately, my data do not perfectly identify acceptance rates nor do I have $t_4 = 1$ and $t_4 = 3$ draft readability scores for every matched pair. Nevertheless, the data I do have and prior research strongly suggest (i)–(iii) not only hold on average, but do not exert upward bias on my estimate of $\underline{D}_{ik}$, more generally. First, I reviewed the literature on gender neutrality in journals' acceptance rates earlier in this section. To recap, women are not accepted more often than men. Results and conclusions are similar when I attempt to adjust for acceptance rates explicitly by also requiring that $T_{5i} \leq T_{5k}$ for matched pairs classified as discrimination against $i$ (Appendix F.6). As shown in Section 3.2, women's draft papers are indeed more readable than men's. Section 3.5 provides further confirmation. Figure 6 plots the readability of women's and men's draft and published papers over increasing $t_4$. Women's drafts are more readable than men's drafts at $t_4 = 3$ *and* their own drafts at $t_4 = 1$.

### *3.4 The time-cost of peer review*

> "Writing simply and directly only looks easy" (Kimble, 1994, p. 53).

Good writing takes time (Hartvigsen, 1981; Kroll, 1990): skilled writers spend longer contemplating a writing assignment, brainstorming and editing; they also write fewer words per minute and produce more drafts (Faigley and Witte, 1981; Stallard, 1974). As a consequence, higher writing standards—and, indeed, higher standards applied more generally (see, *e.g.*, Card et al., 2019; Hengel and Moon, 2019)—should result in female authors spending longer in peer review.

On the other hand, better writing by female economists could perfectly offset some unobservable advantage present in men's papers, conditional on quality. In this case, the time-cost of publishing a

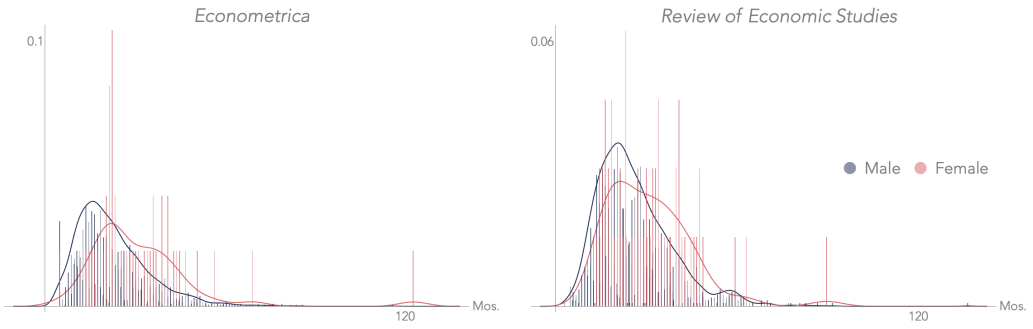


FIGURE 5: Distribution of review times at *Econometrica* and *REStud*

*Notes*. Histograms of time spent under review for papers published in *Econometrica* (left) and *REStud* (right). Blue bars represent papers written only by men (2,398 for *Econometrica*; 1,580 for *REStud*). Pink bars are papers written only by women (48 for *Econometrica*; 64 for *REStud*).

paper will instead be gender neutral—since if it weren't, women could reduce their time spent in review by adopting a strategy marginally closer to men's (or visa versa).[31]

To formalise this idea, consider male and female researchers who use strategies $x_m, x_f \in \mathcal{X}$ to produce papers of identical quality $Q \in \mathcal{Q}$. Let $q$ represent the function mapping $\mathcal{X}$ onto $\mathcal{Q}$ and define $q^{-1}(Q)$, as the set of strategies in $\mathcal{X}$ that achieve the same $Q$.

If men and women are held to identical standards in peer review, then both will accrue identical rewards, conditional on $Q$, *i.e.*,

$$a_m(x_m, Q) = a_f(x_f, Q) = a(x, Q), \tag{14}$$

where $x$ is any strategy in $q^{-1}(Q)$ and $a_g(x_g, Q)$ is the acceptance rate for gender $g \in m, f$ given strategy $x_g$ and quality $Q$.[32] Moreover, if men and women are also equally capable researchers, then neither side should have to exert more effort, conditional on acceptance rate (and, hence, $Q$)—*i.e.*, given Equation (14), there must exist some $\hat{x}_m, \hat{x}_f \in q^{-1}(Q)$ such that

$$c_m(\hat{x}_m) = c_f(\hat{x}_f), \tag{15}$$

where $c_g(x_g)$ is the cost to gender $g$ of implementing the strategy $x_g$.

In the absence of higher standards, Equation (15) implies that men's and women's time-cost of review should be equal, conditional on $Q$. Figure 5 and the analysis in Section 3.4.1 do not support this hypothesis. Figure 5 displays histograms of time (in months) between dates exclusively male- and female-authored papers are first submitted to and their final revisions received by the editorial offices of *Econometria* and *REStud* (mixed-gendered papers are excluded). Women's review times disproportionately cluster above the mean: their articles are five times more likely to experience delays above the 75th percentile than they are to enjoy speedy revisions below the 25th.

**3.4.1 Estimation strategy and results.** For more precision on gender differences in the time-cost of review—and in order to condition explicitly on quality—I build on a model by Ellison (2002):

$$\begin{aligned} \text{revision duration}_j = {} & \beta_0 + \beta_1 \text{ female ratio}_j + \beta_2 \text{ mother}_j + \beta_3 \text{ birth}_j \\ & + \beta_4 \text{ max } t_{5j} + \beta_5 \text{ no. pages}_j + \beta_6 N_j + \beta_7 \text{ order}_j \\ & + \beta_8 \text{ no. citations}_j + \beta_9 \text{ flesch}_j + \boldsymbol{\theta}\, \mathbf{X}_j + \varepsilon_j, \end{aligned} \tag{16}$$

[31] Note that Section 3.3.2 addresses this issue by making a selection-on-observables assumption—*i.e.*, experienced female authors are matched to experienced male authors based on observable characteristics (including citations and field) meant to capture non-readability differences between their papers. One purpose of the present section is to test (indirectly) the robustness of this assumption.

[32] Higher standards come from accepting male-authored papers more often than female-authored papers, conditional on $Q$—*i.e.*, $a_m(x, Q) > a_f(x, Q)$—rewarding men's strategies more than women's strategies even though they both generate identical $Q$—*i.e.*, $a(x_m, Q) > a(x_f, Q)$—or both.

TABLE 8: Revision duration at *Econometrica*, full control set

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Female ratio | 5.379*** | 6.928*** | 6.895*** | 5.811*** | 6.922*** | 9.609*** | 9.595*** |
| | (1.755) | (2.088) | (2.088) | (2.035) | (2.085) | (2.602) | (2.602) |
| Max. $t_5$ | −0.131*** | −0.133*** | −0.132*** | −0.132*** | −0.130*** | −0.126*** | −0.133*** |
| | (0.038) | (0.038) | (0.038) | (0.038) | (0.038) | (0.044) | (0.045) |
| No. pages | 0.196*** | 0.195*** | 0.194*** | 0.195*** | 0.194*** | 0.235*** | 0.222*** |
| | (0.027) | (0.027) | (0.028) | (0.028) | (0.028) | (0.040) | (0.043) |
| $N_j$ | 1.161*** | 1.115*** | 1.102*** | 1.140*** | 1.111*** | 1.433*** | 1.294*** |
| | (0.297) | (0.290) | (0.294) | (0.288) | (0.292) | (0.393) | (0.407) |
| Order | 0.208*** | 0.204*** | 0.203*** | 0.206*** | 0.203*** | 0.466*** | 0.467*** |
| | (0.064) | (0.064) | (0.064) | (0.064) | (0.064) | (0.145) | (0.142) |
| No. citations (asinh) | −0.397** | −0.420** | −0.409** | −0.395* | −0.420** | −0.657 | −0.659* |
| | (0.197) | (0.196) | (0.196) | (0.197) | (0.196) | (0.399) | (0.386) |
| Flesch Reading Ease | −0.018 | −0.016 | −0.016 | −0.018 | −0.016 | −0.034 | −0.037 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.027) | (0.028) |
| Mother | | | −7.866** | | −12.299*** | −22.286*** | −22.331*** |
| | | | (3.498) | | (3.625) | (6.070) | (6.472) |
| Birth | | | | −3.836 | 7.362 | 16.072** | 15.742** |
| | | | | (4.642) | (4.912) | (6.478) | (6.742) |
| Constant | 13.854*** | 13.991*** | 14.006*** | 13.892*** | 14.018*** | 16.239*** | 17.076*** |
| | (1.202) | (1.219) | (1.215) | (1.208) | (1.212) | (2.448) | (2.298) |
| $R^2$ | 0.288 | 0.291 | 0.290 | 0.288 | 0.290 | 0.128 | 0.146 |
| No. observations | 2,622 | 2,607 | 2,622 | 2,622 | 2,622 | 1,278 | 1,278 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Coefficients from OLS estimation of Equation (16); (2) excludes papers authored only by women who gave birth (9 articles) and/or had a child younger than five (15 articles) at some point during peer review; (6) and (7) exclude papers published before 1990. Standard errors clustered by submission year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

where $\text{mother}_j$ and $\text{birth}_j$ are binary variables equal to 1 if article $j$'s authors were all mothers to children younger than five and gave birth, respectively, at some point during peer review,[33] $\max. t_{5j}$ is the number of prior papers published in a top-five economics journal by article $j$'s most prolific co-author, $\text{no. pages}_j$ refers to the page length of the published article, $\text{order}_j$ is the order in which article $j$ appeared in an issue, $\text{no. citations}_j$ are the number of subsequent papers citing $j$, $\text{flesch}_j$ is its Flesch Reading Ease score and $\mathbf{X}_j$ captures additional fixed effects.[34] I first estimate Equation (16) on data from *Econometrica*. I then re-estimate it excluding readability, institution, motherhood and childbirth controls—which I do not have for papers published in *REStud*—on the entire sample and each journal separately.

Table 8 displays results for *Econometrica*. All models include editor, acceptance year and institution fixed effects.[35] Column (1) does not control for motherhood or childbirth; (2) drops papers authored entirely by women who had children younger than five and/or gave birth during peer review; (3) controls for motherhood but not childbirth; (4) controls for childbirth but not motherhood; (5) controls for both

[33] If one co-author goes on maternity leave or has young children, I assume another co-author manages the revision process unless she, too, faces similar family commitments.

[34] Equation (16) controls for all significant factors (plus readability) identified by Ellison (2002). Because authors' English fluency is not significant in his regressions, it is excluded. (Including it has no impact on $\beta_1$.)

[35] See Appendix G.1 for results controlling for years of submission and publication, instead.

TABLE 9: Revision duration at *Econometrica* and *REStud*, restricted control set

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Female ratio | 5.18*** | 1.62 | 3.26*** | 7.70*** | 3.39** | 5.29*** |
| | (1.76) | (1.16) | (1.01) | (2.29) | (1.52) | (1.28) |
| Max. $t_5$ | −0.12*** | −0.11* | −0.12*** | −0.13*** | −0.06 | −0.10** |
| | (0.04) | (0.06) | (0.04) | (0.04) | (0.07) | (0.04) |
| No. pages | 0.20*** | 0.14** | 0.18*** | 0.22*** | 0.06 | 0.17*** |
| | (0.03) | (0.06) | (0.03) | (0.04) | (0.07) | (0.03) |
| $N_j$ | 1.15*** | −0.10 | 0.71** | 1.30*** | 0.18 | 0.99** |
| | (0.30) | (0.49) | (0.27) | (0.41) | (0.66) | (0.37) |
| Order | 0.20*** | −0.08 | 0.08 | 0.47*** | 0.03 | 0.21* |
| | (0.06) | (0.08) | (0.05) | (0.13) | (0.15) | (0.12) |
| No. citations (asinh) | −0.30 | −0.64*** | −0.43*** | −0.43 | −1.20** | −0.84*** |
| | (0.19) | (0.23) | (0.16) | (0.38) | (0.44) | (0.30) |
| Constant | 12.74*** | 23.93*** | 16.96*** | 14.83*** | 30.81*** | 21.64*** |
| | (1.16) | (1.81) | (0.90) | (2.02) | (3.03) | (1.47) |
| $R^2$ | 0.28 | 0.27 | 0.29 | 0.13 | 0.14 | 0.13 |
| No. observations | 2,622 | 1,812 | 4,434 | 1,278 | 1,068 | 2,347 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Coefficients from OLS estimation of Equation (16). Third and sixth column estimates pool data from *Econometrica* and *REStud*; the other four columns were separately estimated on data from each journal. Standard errors clustered by submission year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

childbirth and motherhood; (6) and (7) restrict the sample to papers published after 1990; (7) includes fixed effects for primary *JEL* categories.

Every paper published in *Econometrica* undergoes extensive review, but the consistently large and highly significant coefficient on female ratio suggests women bear the brunt of it. The average male-authored paper takes about 18.5 months to complete all revisions; papers by women need almost seven months longer.

Results pooling data from both journals and on each alone without readability, institution, motherhood and childbirth controls are shown in Table 9. Estimates from *Econometrica* (columns one and four) coincide with those shown in Table 8. Women take 2–3 months longer in review at *REStud* (columns two and five). When observations from both journals are combined, female-authored papers take, on average, 3–5 months longer in peer review (columns three and six).

Remaining coefficients in Table 8 and Table 9 largely correspond to earlier estimates by Ellison (2002). Longer papers take more time to review, as do papers with more co-authors and (generally) those that appear earlier in an issue. Authors with an established publication history, highly cited papers and more readable papers enjoy faster reviews, although the latter effects are only noisily estimated. Giving birth slows down review; having a young child may have the opposite effect.[36]

Appendix G.3 re-estimates column (5) in Table 8 and the third column of Table 9 using a quantile regression model in order to account for a potential rightward-skew in review times (see Figure 5). Appendix G.2 replicates Table 8, column (5) altering the age-threshold on $\text{mother}_j$. The gender gap

[36]This result is consistent with Ginther and Kahn (2004), who find that women with children are more productive than male and childless female doctoral recipients 10 years after receiving their Ph.D. I would interpret it with caution, however, given (i) counter-intuitive results, (ii) obtaining an unbiased estimate of $\beta_2$ was *not* this study's objective and (iii) $\text{mother}_j$ equals one for only a small number of articles in the sample.

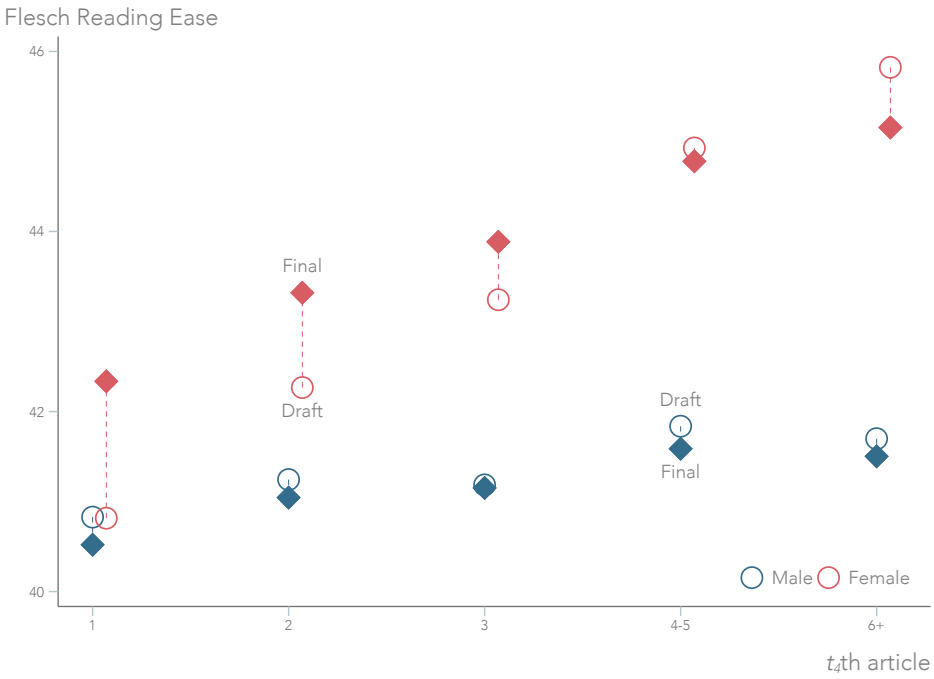


FIGURE 6: Readability of authors' $t_4$th top-four publication (draft and final)

*Notes.* Flesch Reading Ease marginal mean scores for male (blue) and female (pink) authors' $t_4 = 1$, $t_4 = 2$, *etc.* top-four publications. Hollow circles denote draft readability; solid diamonds denote readability in published versions of the same papers. See Table 10 for point estimates, standard errors and further estimation details.

is not driven by outliers—the coefficient on female ratio is positive and significant across the entire distribution—nor does it depend on the precise definition of motherhood.

### *3.5 Understanding how women respond to higher standards*

Women can respond to higher standards in two different ways: immediately (direct effect) and pre-emptively (feedback effect). As emphasised in Section 3.3.1, the weight of each effect likely depends on authors' information about—hence experience with—the peer review process.

To illustrate the evolution of the relative importance of each, I compare papers pre- and post-review as authors' publication counts rise (Equation (17)):

$$R_{j_{it}m} = \beta_0 + \beta_1 \text{ female ratio}_j + \beta_2 \text{ female ratio}_j \times t_{4it} + \beta_3 t_{4it} + \boldsymbol{\theta}\, \mathbf{X}_j + \varepsilon_j, \tag{17}$$

where $m = W, P$ for working papers and published articles, respectively, $t_{4it}$ is author $i$'s number of top-four papers at time $t$ and $\mathbf{X}_j$ is a vector of observable controls.[37]

**3.5.1 Results.** Results are shown in Figure 6 and Table 10. In Figure 6, hollow circles denote draft readability; solid diamonds reflect readability in the final, published versions of those same papers. Dashed lines trace readability as papers undergo peer review (direct effect) and correspond to estimates in the first panel of Table 10. Table 10's second panel shows the effect of female ratio ($\beta_2$) for each version of a manuscript. Figures in the final row represent gender differences in the direct effect.[38]

Figure 6 and Table 10 suggest that gender differences in the direct effect of peer review start off large, positive and significant; as $t_4$ increases, they gradually go away. For the feedback effect, however,

[37] As in Section 3.3.2, data are disaggregated to the author-level by duplicating each article $N_j$ times; to account for duplicate articles, regressions are weighted by $1/N_j$. Results and conclusions based on unweighted regressions—or by replacing $t_{4it}$ with max. $t_{4j}$ and *not* duplicating articles—are very similar to those presented here.

[38] The difference between final row estimates at $t_4 = 1$ and $t_4 = 3$ is weakly statistically significant (standard error 0.66).

TABLE 10: Readability of authors' $t_4$th top-four publication (draft and final)

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ |
|---|---|---|---|---|---|
| **Predicted $R_{jP} - R_{jW}$** | | | | | |
| Women | 1.52** | 1.05* | 0.64 | −0.15 | −0.67 |
| | (0.64) | (0.61) | (0.71) | (0.91) | (1.19) |
| Men | −0.31* | −0.20* | −0.03 | −0.25 | −0.20 |
| | (0.18) | (0.11) | (0.10) | (0.16) | (0.19) |
| **Marginal effect of female ratio** | | | | | |
| Published article | 1.82* | 2.27*** | 2.73*** | 3.19*** | 3.65** |
| | (1.02) | (0.74) | (0.76) | (1.07) | (1.50) |
| Draft paper | −0.01 | 1.02 | 2.06** | 3.09*** | 4.12*** |
| | (1.22) | (0.94) | (0.84) | (0.97) | (1.27) |
| **Difference** | 1.83*** | 1.25* | 0.68 | 0.10 | −0.47 |
| | (0.70) | (0.67) | (0.79) | (1.01) | (1.28) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 4,289 observations; 1,988 and 1,986 distinct NBER working papers and published articles, respectively; 1,839 distinct authors. Panel one displays magnitude of predicted $R_{jP} - R_{jW}$ (the direct effect of peer review) for women and men over increasing publication counts ($t_4$). Panel two estimates the marginal effect of an article's female ratio ($\beta_1 + \beta_2 \times t_4$), separately for draft papers and published articles. Figures from FGLS estimation of Equation (17). Quality controls denoted by ✓[2] include citation count (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority). Standard errors clustered by editor and robust to cross-model correlation in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

the pattern is reversed: although the readability gap in published articles is statistically significant and relatively stable at every $t_4$, it increasingly forms before submission. Differences in draft readability contribute nothing to the gap at $t_4 = 1$. That rises to 40 percent at $t_4 = 2$ and 70 percent at $t_4 = 3$. By $t_4 = 4$–$5$ and $t_4 = 6+$, both sexes largely address referee concerns before peer review.

**3.5.2 Interpretation.** A number of *tentative* conclusions about the gender readability gap—and informative about gender differences in preferences and beliefs—can be made from Figure 6 and Table 10.

First, inexperienced men and women seem to make similar choices in draft readability. This suggests identical initial preferences for and beliefs about the impact of writing well. In one important sense, however, men are still better informed: the standards they believe apply actually do; junior women appear to mistakenly assume similar standards apply to them, too.

Second, experienced men *and* women seem to sacrifice time upfront in order to improve their odds in peer review. By anticipating referees' demands, authors can partially insure themselves against rejection and/or excessively long review. The price is having to spend more time revising a manuscript before submitting it. Assuming choices by senior economists express optimal trade-offs with full information, Figure 6 implies little—if any—gender differences in these insurance preferences. Nevertheless, higher standards mean the price of that insurance is greater for women than it is for men.

Finally, Figure 6 suggests the direct effect of peer review dominates when women have less experience; the feedback effect dominates when women have more experience. This pattern of behaviour implies that women initially underestimate referees' thresholds but learn about them over time and adapt their *ex ante* writing style accordingly.

This last observation suggests inexperienced female economists go through the toughest review, conditional on acceptance. To investigate further, I test the impact of experience on time spent in

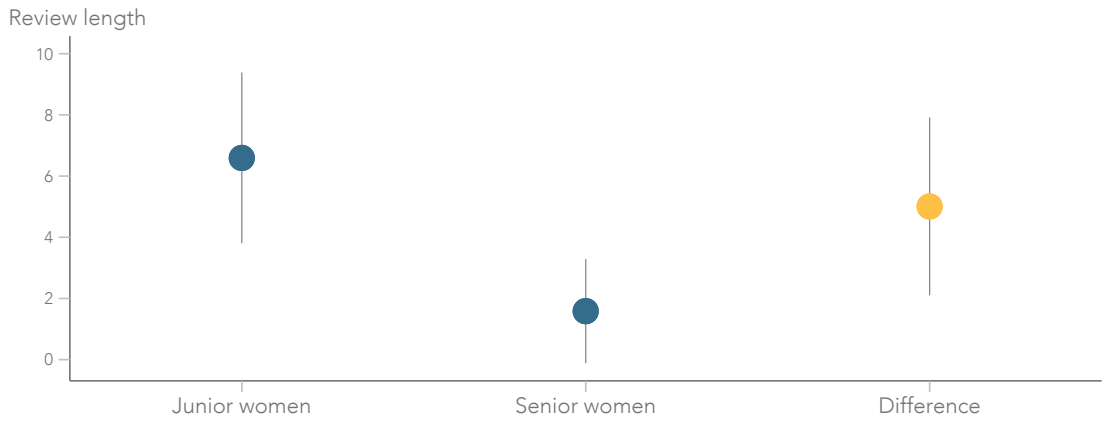


FIGURE 7: The effect of experience on women's review times

*Notes*. Sample 5,019 observations satisfying max. $t_{5j} = t_{5i}$; 4,436 distinct articles and 2,429 distinct authors (166 female). Blue dots are the coefficients on female ratio corresponding to separate FGLS estimations of Equation (11) on authors for whom $t_{5i} = 1$ (junior) and $t_{5i} > 1$ (senior), respectively. The yellow dot is their difference. Vertical grey lines correspond to 90 percent confidence intervals. Regressions weighted by $1/N_j$.

review by re-estimating Equation (16) on sub-samples of junior ($t_5 = 1$) and senior ($t_5 > 1$) authors.[39] Results are displayed in Figure 7. They suggest papers by junior women do indeed take longer in review; the gender gap is significantly smaller—albeit still positive—for senior women.

## 4 Conclusion

Most raw numerical counts suggest women produce less than men: female real estate agents list fewer homes (Seagraves and Gallimore, 2013); female lawyers bill fewer hours (Azmat and Ferrer, 2017); female physicians see fewer patients (Bloor et al., 2008); female academics write fewer papers (Ceci et al., 2014). When evaluated by narrowly defined quality measures, however, women often outperform: houses listed by female real estate agents sell for higher prices (Salter et al., 2012; Seagraves and Gallimore, 2013); female lawyers make fewer ethical violations (Hatamyar and Simmons, 2004); patients treated by female physicians are less likely to die or be readmitted to hospital (Tsugawa et al., 2017).

As I show in this paper, female economists surpass men on another dimension: writing clarity. Using five well-known readability scores, I analyse every article abstract published in a top four economics journal since 1950. Abstracts written by women are 1–6 percent more readable. A comparison of published papers to their pre-reviewed drafts suggests the immediate impact of peer review directly explains at least forty percent of this gap.

Why? Either women voluntarily improve their writing during peer review—*e.g.*, because they're more sensitive to criticism—or better written papers are women's response to higher standards imposed by referees and/or editors. To theoretically distinguish between hypotheses, I construct a dynamic model of an author's decision-making process. To empirically test it, I exploit within- and between-individual readability changes among well-published economists. My results suggest higher standards cause experienced women to write at least 5–7 percent more clearly than they otherwise would.

Higher standards hurt women's productivity and labour market outcomes. Work that is evaluated more critically *at any point in the production process* will be systematically better (holding prices fixed) or

[39] Three notes on estimation. First, in Section 3.3.2, I define "experienced" as $t_4 = 3$. However, most female-authored papers published in *Econometrica* and *REStud* are by women with no (or only one) previous top publication; only 24 have two or more previous papers and were the most senior co-author on a $t_4 > 2$ paper. Second, to eliminate confounding by more senior co-authors, I restrict the sample to authors satisfying max. $t_{5j} = t_{5i}$. (Including these observations does not substantially impact results or conclusions). Third, because the sample includes data from *REStud*, readability, institution, motherhood and childbirth controls are not included. See the August 2018 version of this paper for results that control for these factors (based on data from *Econometrica* alone).

systematically cheaper (holding quality fixed). This will reduce women's wages and distort measurement of their productivity. For example, if judges require better writing in female-authored briefs, female attorneys must charge lower fees and/or under-report hours to compete with men; billable hours and client revenue will decline, making female lawyers appear less productive than they truly are.

Unfortunately, there are no easy solutions for addressing higher standards. But least intrusive—and arguably most effective—is simple awareness and constant supervision. I hope journals are challenged to address the tougher standards they likely impose on women, open to policies that transparently monitor them and supportive of research that helps us better understand them.

## References


Abrevaya, J. and D. S. Hamermesh (2012). "Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)?" *Review of Economics and Statistics* 94 (1), pp. 202–207 (cit. on p. 3).

Alexander, D., O. Gorelkina, and E. Hengel (2018). "A reply to Tol, 2018". Mimeo (cit. on p. 3).

Altonji, J. G. and C. R. Pierret (2001). "Employer Learning and Statistical Discrimination". *Quarterly Journal of Economics* 116 (1), pp. 313–350 (cit. on p. 4).

Antecol, H., K. Bedard, and J. Stearns (Sept. 2018). "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" *American Economic Review* 108 (9), pp. 2420–2441 (cit. on p. 4).

Anwar, S. and H. Fang (2006). "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence". *American Economic Review* 96 (1), pp. 127–151 (cit. on p. 4).

— (2015). "Testing for Racial Prejudice in the Parole Board Release Process: Theory and Evidence". *The Journal of Legal Studies* 44 (1), pp. 1–37 (cit. on p. 4).

Ashenfelter, O. and A. Krueger (1994). "Estimates of the Economic Return to Schooling from a New Sample of Twins". *American Economic Review* 84 (5), pp. 1157–1173 (cit. on p. 9).

Azmat, G. and R. Ferrer (2017). "Gender Gaps in Performance: Evidence from Young Lawyers". *Journal of Political Economy* 125 (5), pp. 1306–1355 (cit. on p. 27).

Babcock, L. and S. Laschever (2003). *Women Don't Ask: Negotiation and the Gender Divide*. Princeton, New Jersey: Princeton University Press (cit. on p. 4).

Babcock, L. et al. (2017). "Gender differences in accepting and receiving requests for tasks with low promotability". *American Economic Review* 107 (3), pp. 714–747 (cit. on p. 2).

Bayer, A. and C. E. Rouse (2016). "Diversity in the Economics Profession: A New Attack on an Old Problem". *Journal of Economic Perspectives* 30 (4), pp. 221–242 (cit. on p. 3).

Bertrand, M., C. Goldin, and L. F. Katz (2010). "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors". *American Economic Journal: Applied Economics* 2 (3), pp. 228–255 (cit. on p. 3).

Bertrand, M. and S. Mullainathan (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". *American Economic Review* 94 (4), pp. 991–1013 (cit. on p. 4).

Blank, R. M. (1991). "The Effects of Double-blind versus Single-blind Reviewing: Experimental Evidence from the American Economic Review". *American Economic Review* 81 (5), pp. 1041–1067 (cit. on pp. 1, 12, 17).

Blau, F. D. and L. M. Kahn (2017). "The Gender Wage Gap: Extent, Trends, and Explanations". *Journal of Economic Literature* 55 (3), pp. 789–865 (cit. on pp. 3, 4).

Bloor, K., N. Freemantle, and A. Maynard (2008). "Gender and Variation in Activity Rates of Hospital Consultants". *Journal of the Royal Society of Medicine* 101 (1), pp. 27–33 (cit. on p. 27).

Bordalo, P. et al. (2016). "Stereotypes". *Quarterly Journal of Economics* 131 (4), pp. 1753–1794 (cit. on p. 4).

Boring, A. (2017). "Gender Biases in Student Evaluations of Teaching". *Journal of Public Economics* 145 (Supplement C), pp. 27–41 (cit. on p. 3).

Born, A., E. Ranehill, and A. Sandberg (2019). "A man's world? – The Impact of a Male-Dominated Environment on Female Leadership". Mimeo (cit. on p. 2).
Borsuk, R. M. et al. (2009). "To Name or Not to Name: The Effect of Changing Author Gender on Peer Review". *BioScience* 59 (11), pp. 985–989 (cit. on p. 1).
Bransch, F. and M. Kvasnicka (2017). "Male Gatekeepers Gender Bias in the Publishing Process?" IZA Discussion Paper Series, No. 11089 (cit. on p. 3).
Bright, L. K. (2017). "Decision Theoretic Model of the Productivity Gap". *Erkenntnis* 82 (2), pp. 421–442 (cit. on p. 3).
Card, D. and S. DellaVigna (2013). "Nine Facts about Top Journals in Economics". *Journal of Economic Literature* 51 (1), pp. 144–161 (cit. on p. 3).
— (2017). "What do Editors Maximize? Evidence from Four Leading Economics Journals". NBER Working Paper Series, No. 23282 (cit. on p. 3).
Card, D. et al. (2019). "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics* 135 (1), pp. 269–327 (cit. on pp. 2, 3, 17, 21).
Carlana, M. (2019). "Implicit Stereotypes: Evidence from Teachers' Gender Bias". *Quarterly Journal of Economics* 134 (3), pp. 1163–1224 (cit. on p. 4).
Casnici, N. et al. (2016). "Assessing peer review by guaging the fate of rejected manuscripts . The case of Journal of Artificial Societies and Social Simulation Why looking at the fate of unpublished manuscripts ?" *PEERE Meeting: Vaxjo 2016* (September) (cit. on p. 3).
Ceci, S. J. et al. (2014). "Women in Academic Science: A Changing Landscape". *Psychological Science in the Public Interest* 15 (3), pp. 75–141 (cit. on pp. 17, 27).
Chari, A. and P. Goldsmith-Pinkham (2017). "Gender Representation in Economics Across Topics and Time: Evidence from the NBER Summer Institute". NBER Working Paper Series, No. 23953 (cit. on pp. 3, 17).
Clain, S. H. and K. Leppel (2018). "Patterns in Economics Journal Acceptances and Rejections". *American Economist* 63 (1), pp. 94–109 (cit. on p. 3).
Coate, S. and G. C. Loury (1993). "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5), pp. 1220–1240 (cit. on p. 4).
Coffman, K. B. (2014). "Evidence on Self-stereotyping and the Contribution of Ideas". *Quarterly Journal of Economics* 129 (4), pp. 1625–1660 (cit. on pp. 2, 4).
Correll, S. and C. Simard (2016). "Vague Feedback Is Holding Women Back". *Harvard Business Review*. https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back. Accessed: 2016-10-04 (cit. on p. 3).
Cortés, P. and J. Pan (2016). "Prevalence of Long Hours and Women's Job Choices: Evidence across Countries and within the U.S." IZA Discussion Paper Series, No. 10225 (cit. on p. 4).
Croson, R. and U. Gneezy (2009). "Gender Differences in Preferences". *Journal of Economic Literature* 47 (2), pp. 448–474 (cit. on p. 2).
Ductor, L., S. Goyal, and A. Prummer (2018). "Gender and Collaboration". Mimeo. Cambridge (cit. on p. 3).
Ellison, G. (2002). "The Slowdown of the Economics Publishing Process". *Journal of Political Economy* 110 (5), pp. 947–993 (cit. on pp. 3, 10, 22–24).
Exley, C. L. and J. B. Kessler (2019). "The Gender Gap in Self-Promotion". Mimeo (cit. on p. 2).
Faigley, L. and S. P. Witte (1981). "Analyzing Revision". *College Composition and Communication* 32 (4), pp. 400–414 (cit. on p. 21).
Fang, F. C., J. W. Bennett, and A. Casadevall (2013). "Males Are Overrepresented among Life Science Researchers Committing Scientific Misconduct". *mBio* 4 (1), pp. 1–3 (cit. on p. 4).
Foschi, M. (1996). "Double Standards in the Evaluation of Men and Women". *Social Psychology Quarterly* 59 (3), pp. 237–254 (cit. on p. 1).
Fryer, R. G., D. Pager, and J. L. Spenkuch (2013). "Racial Disparities in Job Finding and Offered Wages". *Journal of Law and Economics* 56 (3), pp. 633–689 (cit. on p. 4).

Gans, J. S. and G. B. Shepherd (1994). "How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists". *Journal of Economic Perspectives* 8 (1), pp. 165–179 (cit. on p. 1).
Gardiner, B. et al. (2016). "The Dark Side of Guardian Comments". *Guardian*. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments. Accessed: 2016-10-04 (cit. on p. 3).
Gilbert, J. R., E. S. Williams, and G. D. Lundberg (1994). "Is There Gender Bias in JAMA's Peer Review Process?" *Journal of the American Medical Association* 272 (2), pp. 139–142 (cit. on pp. 1, 17).
Ginther, D. K. and S. Kahn (2004). "Women in Economics: Moving Up or Falling Off the Academic Career Ladder?" *Journal of Economic Perspectives* 18 (3), pp. 193–214 (cit. on pp. 3, 24).
Glover, D., A. Pallais, and W. Pariente (2017). "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores". *Quarterly Journal of Economics* 132 (3), pp. 1219–1260 (cit. on p. 4).
Goldberg, P. (1968). "Are Women Prejudiced against Women?" *Trans-action* 5 (5), pp. 28–30 (cit. on p. 1).
Goldberg, P. K. (2014). "Report of the Editor: American Economic Review". *American Economic Review* 104 (5), pp. 621–631 (cit. on p. 12).
— (2015). "Report of the Editor: American Economic Review". *American Economic Review* 105 (5), pp. 698–710 (cit. on p. 10).
Goldin, C. (2014a). "A Grand Gender Convergence: Its Last Chapter". *American Economic Review* 104 (4), pp. 1091–1119 (cit. on p. 3).
— (2014b). "A Pollution Theory of Discrimination: Male and Female Differences in Occupations and Earnings". In: *Human Capital in History: The American Record*. Ed. by L. P. Boustan, C. Frydman, and R. A. Margo. Cambridge, Massachusetts: National Bureau of Economic Research, pp. 313–348 (cit. on p. 4).
Goldin, C. and L. F. Katz (2016). "A Most Egalitarian Profession: Pharmacy and the Evolution of a Family-Friendly Occupation". *Journal of Labor Economics* 34 (3), pp. 705–746 (cit. on p. 3).
Goldin, C. and C. Rouse (2000). "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians". *American Economic Review* 90 (4), pp. 715–741 (cit. on p. 3).
Grossbard, S., T. Yilmazer, and L. Zhang (2018). "The Gender Gap in Citations of Economics Articles: Lessons from Economics of the Household". Mimeo (cit. on p. 3).
Grunspan, D. Z. et al. (2016). "Males Under-estimate Academic Performance of Their Female Peers in Undergraduate Biology Classrooms". *PLOS ONE* 11 (2), pp. 1–16 (cit. on p. 1).
Handley, G. et al. (2015). "An Examination of Gender Differences in the American Fisheries Society Peer-Review Process". *Fisheries Magazine* 40 (9), pp. 442–451 (cit. on p. 17).
Hart, R. L. (2000). "Co-authorship in the Academic Library Literature: A Survey of Attitudes and Behaviors". *Journal of Academic Librarianship* 26 (5), pp. 339–345 (cit. on p. 6).
Hartley, J., J. W. Pennebaker, and C. Fox (2003). "Using New Technology to Assess the Academic Writing Styles of Male and Female Pairs and Individuals". *Journal of Technical Writing and Communication* 33 (3), pp. 243–261 (cit. on p. 4).
Hartvigsen, M. K. (1981). "A Comparative Study of Quality and Syntactic Maturity between In-class and Out-of-class Writing Samples of Freshmen at Washington State University". PhD thesis. Washington State University (cit. on p. 21).
Hatamyar, P. W. and K. M. Simmons (2004). "Are Women More Ethical Lawyers? An Empirical Study". *Florida State University Law Review* 31 (4), pp. 785–858 (cit. on pp. 4, 27).
Heilman, M. E. and M. C. Haynes (2005). "No Credit Where Credit Is Due: Attributional Rationalization of Women's Success in Male-female Teams". *Journal of Applied Psychology* 90 (5), pp. 905–916 (cit. on p. 1).
Hengel, E. (2016). "Publishing while Female: Gender Differences in Peer Review Scrutiny". Mimeo (cit. on p. 21).
— (2017). "Publishing while Female: Are Women Held to Higher Standards? Evidence From Peer Review." Cambridge Working Paper Economics: 1753 (cit. on pp. 4, 20, 21).

Hengel, E. and E. Moon (2019). "Gender and Quality at Top Economics Journals". Mimeo (cit. on pp. 3, 21).
Hengel, E. and R. Tol (2018). "Gender and the Review Process at 35 Economics and Finance Journals". Mimeo (cit. on p. 3).
Jenkins, S. (2007). "A Woman's Work Is Never Done? Fund-Raising Perception and Effort among Female State Legislative Candidates". *Political Research Quarterly* 60 (2), pp. 230–239 (cit. on p. 4).
Kimble, J. (1994). "Answering the Critics of Plain Language". *Scribes Journal of Legal Writing* 51 (1994-1995), pp. 51–85 (cit. on p. 21).
Knowles, J., N. Persico, and P. Todd (2001). "Racial Bias in Motor Vehicle Searches: Theory and Evidence". *Journal of Political Economy* 109 (1), pp. 203–229 (cit. on p. 4).
Krawczyk, M. and M. Smyk (2016). "Author's Gender Affects Rating of Academic Articles: Evidence from an Incentivized, Deception-free Laboratory Experiment". *European Economic Review* 90, pp. 326–335 (cit. on p. 1).
Kroll, B. (1990). "What Does Time Buy? ESL Student Performance on Home versus Class Compositions". In: *Second Language Writing*. Ed. by B. Kroll. Cambridge, U.K.: Cambridge University Press. Chap. 9, pp. 140–154 (cit. on p. 21).
Kumar, S. and K. Ratnavelu (2016). "Perceptions of Scholars in the Field of Economics on Coauthorship Associations: Evidence from an International Survey". *PLoS ONE* 11 (6), pp. 1–18 (cit. on p. 6).
Lavy, V. and E. Sand (2015). "On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases". NBER Working Paper Series, No. 20909 (cit. on p. 4).
Lloyd, M. E. (1990). "Gender Factors in Reviewer Recommendations for Manuscript Publication". *Journal of Applied Behavior Analysis* 23 (4), pp. 539–543 (cit. on p. 1).
Lundberg, S. J. and R. Startz (1983). "Private Discrimination and Social Intervention in Competitive Labor Markets". *American Economic Review* 73 (3), pp. 340–347 (cit. on p. 4).
McGillivray, B. and E. De Ranieri (2018). "Uptake and Outcome of Manuscripts in Nature Journals by Review Model and Author Characteristics". Mimeo (cit. on p. 17).
Mengel, F., J. Sauermann, and U. Zölitz (2017). "Gender Bias in Teaching Evaluations". 00 (April), pp. 1–45 (cit. on p. 3).
Möbius, M. M. et al. (2014). "Managing Self-Confidence". Mimeo (cit. on p. 2).
Mohr, T. S. (2014). "Why Women Don't Apply for Jobs Unless They're 100% Qualified". *Harvard Business Review*. https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified. Accessed: 2017-11-16 (cit. on p. 4).
Moss-Racusin, C. A. et al. (2012). "Science Faculty's Subtle Gender Biases Favor Male Students". *Proceedings of the National Academy of Sciences* 109 (41), pp. 16474–16479 (cit. on p. 1).
Neumark, D., R. J. Bank, and K. D. Van Nort (1996). "Sex Discrimination in Restaurant Hiring: An Audit Study". *Quarterly Journal of Economics* 111 (3), pp. 915–941 (cit. on p. 4).
Niederle, M. and L. Vesterlund (2010). "Explaining the Gender Gap in Math Test Scores: The Role of Competition". *Journal of Economic Perspectives* 24 (2), pp. 129–144 (cit. on p. 3).
Paludi, M. A. and W. D. Bauer (1983). "Goldberg Revisited: What's in an Author's Name". *Sex Roles* 9 (3), pp. 387–390 (cit. on p. 1).
Parsons, C. A. et al. (2011). "Strike Three: Discrimination, Incentives, and Evaluation". *American Economic Review* 101 (4), pp. 1410–1435 (cit. on p. 4).
Plavén-Sigray, P. et al. (2017). "The Readability of Scientific Texts is Decreasing over Time". *eLife* 6 (e27725), pp. 1–14 (cit. on p. 4).
Reuben, E., P. Sapienza, and L. Zingales (2014). "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences* 111 (12), pp. 4403–4408 (cit. on p. 1).
Roter, D. L. and J. A. Hall (2004). "Physician Gender and Patient-centered Communication: A Critical Review of Empirical Research". *Annual Review of Public Health* 25 (May), pp. 497–519 (cit. on p. 4).

Salter, S. P. et al. (2012). "Broker Beauty and Boon: A Study of Physical Attractiveness and Its Effect on Real Estate Brokers' Income and Productivity". *Applied Financial Economics* 22 (10), pp. 811–825 (cit. on p. 27).
Sarsons, H. (2019). "Gender Differences in Recognition for Group Work". *Journa of Political Economy* (forthcoming) (cit. on pp. 1, 3).
Schlosser, A., Z. Neeman, and Y. Attali (2019). "Differential Performance in High Versus Low Stakes Tests: Evidence from the Gre Test". *Economic Journal* 129 (10), pp. 2916–2948 (cit. on p. 2).
Seagraves, P. and P. Gallimore (2013). "The Gender Gap in Real Estate Sales: Negotiation Skill or Agent Selection?" *Real Estate Economics* 41 (3), pp. 600–631 (cit. on p. 27).
Stallard, C. K. (1974). "An Analysis of the Writing Behavior of Good Student Writers". *Research in the Teaching of English* 8 (2), pp. 206–218 (cit. on p. 21).
Teele, D. L. and K. Thelen (2017). "Gender in the Journals: Publication Patterns in Political Science". *PS - Political Science and Politics* 50 (2), pp. 433–447 (cit. on p. 3).
Tol, R. S. (2018). "Gender at Energy Economics". *Energy Economics* 72, pp. 558–559 (cit. on p. 17).
Tsugawa, Y. et al. (2017). "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians". *JAMA Internal Medicine* 177 (2), pp. 206–213 (cit. on p. 27).
Wu, A. H. (2019). "Gender Bias in Rumors among Professionals: An Identity-based Interpretation". *Review of Economics and Statistics* (forthcoming) (cit. on p. 3).

# Appendices

## A Proofs for Theorem 1 and Corollary 1

The proof of Theorem 1 follows directly from Lemma 5, at the end of this section. The proof of Lemma 5 relies on a series of additional lemmas stated and proved below. It is followed by a proof of Corollary 1. Throughout,

- $\{(r_{0it}, R_{it})\}$ represents the sequence of readability choices made by author $i$ for all $t$;
- Without loss of generality, $c_{it}(0)$ and $\phi_i(0)$ are 0;
- $R_i^\star$ is defined as the $R$ that solves $\phi_i'(R) = c_i'(R)$;
- $t > \underline{t}$ where $\underline{t}$ is large enough that $\left|c_{i\underline{t}}(r) - c_i(r)\right|$ is inconsequential for all $r \in \mathbb{R}$;[1]
- Review group $s$ is referred to as "state $s$".

**Lemma 1.** *$\{(r_{0it}, R_{it})\}$ is bounded.*

*Proof.* Consider the sequence of initial readability choices, $\{r_{0it}\}$. I first show that $R_i^\star \leq r_{0it}$ for all $t$. Recall $r_{0it}$ is chosen to maximise the author's subjective expected utility in Equation (9). It satisfies the following first order condition

$$\int_\Sigma \left( \pi_{0it}^s(r_{0it}) v_{1it}^s + \Pi_{0it}^s(r_{0it}) \frac{\partial v_{1it}^s}{\partial r_{0it}} \right) \mathrm{d}\mu_i + \phi_i'(r_{0it}) - c_i'(r_{0it}) = 0, \tag{A.1}$$

where $v_{1it}^s$ represents Equation (9) evaluated at the optimal $r_{1it}$.

$\phi_{i|r_{0it}}(r_{1it}) = \phi_i(R_{it}) - \phi_i(r_{1it})$ and $c_{i|r_{0it}}(r_{1it}) = c_i(R_{it}) - c_i(r_{0it})$. Thus,

$$\begin{aligned} \frac{\partial v_{1it}^s}{\partial r_{0it}} &= \pi_{1it}^s(R_{it}) u_i + \phi_i'(R_{it}) - c_i'(R_{it}) - \phi_i'(r_{0it}) + c_i'(r_{0it}) \\ &= \frac{\partial v_{1it}^s}{\partial r_{1it}} + c_i'(r_{0it}) - \phi_i'(r_{0it}). \end{aligned} \tag{A.2}$$

Since $\phi_i'(R_i^\star) = c_i'(R_i^\star)$, $\partial v_{1it}^s / \partial r_{0it} = \partial v_{1it}^s / \partial r_{1it}$ when evaluated at $r_{0it} = R_i^\star$. The left hand side of Equation (A.1) evaluated at $r_{0it} = R_i^\star$ is correspondingly equivalent to

$$\int_\Sigma \left( \pi_{0it}^s(r_{0it}) v_{1it}^s + \Pi_{0it}^s(r_{0it}) \frac{\partial v_{1it}^s}{\partial r_{1it}} \right) \mathrm{d}\mu_i. \tag{A.3}$$

$v_{1it}^s$ is non-negative;[2] optimising behaviour at stage 1 implies $\partial v_{1it}^s / \partial r_{1it} \geq 0$: either an $r_{1it}$ exists that satisfies $\partial v_{1it}^s / \partial r_{1it} = 0$, or the author chooses $r_{1it} = 0$ and $\partial v_{1it}^s / \partial r_{1it} = \pi_{1it}^s(R_{it}) u_i$ is non-negative. Thus, Equation (A.3) is non-negative. Since $c_i'(r) < \phi_i'(r)$ for all $r < R_i^\star$, the left-hand side of Equation (A.1) is strictly positive for all $r < R_i^\star$, so $r_{0it}$ must be at least as large as $R_i^\star$.

I now show that $\{r_{0it}\}$ is bounded from above. As $r_0$ tends to infinity, authors choose not to make any changes at stage 1. Thus,

$$\lim_{r_0 \to \infty} \Pi_{0it}^s(r_0) v_{1it}^s = \overline{\Pi}_{0it}^s \overline{\Pi}_{1it}^s u_i, \tag{A.4}$$

where $\overline{\Pi}_{0it}^s$ and $\overline{\Pi}_{1it}^s$ are some upper bounds on the author's subjective probability of receiving an R&R and then being accepted in state $s$ at time $t$. Since both are no more than 1, $u_i$ is finite and $\phi_i(r) - c_i(r)$ is strictly decreasing for all $r > R_i^\star$,

$$\lim_{r_0 \to \infty} \left\{ \int_\Sigma \Pi_{0it}^s(r_0) v_{1it}^s \, \mathrm{d}\mu_i + \phi_i(r_0) - c_i(r_0) \right\} = -\infty. \tag{A.5}$$

[1] That is, I assume throughout the proof that $t$ is large enough for $c_i$ to be a sufficiently close approximation of $c_{it}$.

[2] Equation (8) evaluated at $r_{1it} = 0$ is non-negative. Since $r_{1it}$ maximises Equation (8), $v_{1it}^s$ is likewise non-negative.

Similarly, because $\Pi^s_{0it}(r_{0it})\Pi^s_{1it}(R_{it}) \leq 1$ for all $s$ and $\phi_i(r)$ and $c_i(r)$ are finite at all $r < \infty$, Equation (9) is likewise finite for all $r < \infty$. Thus,

$$\sup\left\{\underset{r_{0it}}{\operatorname{argmax}} \int_\Sigma \Pi^s_{0it}(r_{0it})v^s_{1it}\,\mathrm{d}\mu_i + \phi_i(r_{0it}) - c_i(r_{0it})\right\} < \infty,$$

so $\{r_{0it}\}$ is bounded.

It remains to show that $\{R_{it}\}$ is likewise bounded. Since $r_{1it} \geq 0$ and $R_{it} = r_{0it} + r_{1it}$, $R_{it}$ is bounded below by $r_{0it}$, which, as just shown, is itself bounded. Additionally, the author opts for $r_{1it} = 0$ if Equation (8) is less than 0 for all $r_{1it} > 0$. Since $R^\star_i \leq r_{0it}$ and $\Pi^s_{1it}(R_{it}) \leq 1$

$$\begin{aligned}&\Pi^s_{1it}(R_{it})u_i + \phi_i(R_{it}) - \phi_i(r_{0it}) - c_i(R_{it}) + c_i(r_{0it})\\ &\quad\leq u_i + \phi_i(R_{it}) - c_i(R_{it}).\end{aligned} \tag{A.6}$$

Equation (A.6) is strictly decreasing in $R$ for all $R \geq R^\star_i$. The author will not choose any $R$ strictly greater than the one that equates Equation (A.6) to 0. Thus, $\{R_{it}\}$ is bounded from above.

Because $\{r_{0it}\}$ and $\{R_{it}\}$ are bounded, the sequence $\{(r_{0it}, R_{it})\}$ in $\mathbb{R}^2$ is likewise bounded. Thus, all is proved. □

**Lemma 2.** *$r_{0i} \leq r_{0it}$ and $R^s_i \leq R^s_{it}$ for all $t > t''$.*

*Proof.* Bounded infinite sequences have at least one cluster point and at least one subsequence that converges to each cluster point (Bolzano-Weierstrass). Let $\{(r_{0it}, R^{q^\star}_{it})\}$ denote the complete subsequence of $\{(r_{0it}, R_{it})\}$ in which state $q$ is reached. Thus,

$$\left\{\left(r_{0it}, R^{s^\star}_{it}\right)\right\} \bigcap_{s^\star \neq q^\star} \left\{\left(r_{0it}, R^{q^\star}_{it}\right)\right\} = \varnothing \quad \text{and} \quad \bigcup_{q^\star \in \Sigma} \left\{\left(r_{0it}, R^{q^\star}_{it}\right)\right\} = \{(r_{0it}, R_{it})\}.$$

Fix state $s$. Because $\Sigma$ is finite, $\{(r_{0it}, R^{s^\star}_{it})\}$ likewise forms a bounded infinite sequence and therefore converges to at least one cluster point. Fix one such cluster point, $(r_{0i}, R^s_i)$, and let $\{(r_{0it}, R^s_{it})\}$ denote the subsequence of $\{(r_{0it}, R^{s^\star}_{it})\}$ that converges to it.

Consider first the proposition that $R^s_i \leq R^s_{it}$ for all $t > t''$. By way of a contradiction, assume $R^s_{it} < R^s_i$ for all $t > t''$ and some fixed $r^s_{0it}$. Thus, $r^s_{1it} < r^s_{1it+1}$ for all $t > t''$. A positive $r^s_{1it}$ implies that $R^s_{it}$ satisfies

$$\pi^s_{1it}(R^s_{it}) = \frac{1}{u_i}\left(c'_i(R^s_{it}) - \phi'_i(R^s_{it})\right). \tag{A.7}$$

Let $\pi^s_{1i}$ denote the terminal value of $\pi^s_{1it}$ as $t$ tends to $\infty$. $\pi^s_{1i}$ is finite; thus, $\{\pi^s_{1it}\}$ itself converges: if $\widetilde{R}^s_i < R^s_i$, then $\pi^s_{1it}(R^s_{it}) = 0$ for all $t > t''$, where $t''$ has been redefined to assure $\widetilde{R}^s_i \leq R^s_{it}$; if $R^s_i \leq \widetilde{R}^s_i$ and $\pi^s_{1i}(R^s_i) = \infty$, then $\pi^s_{1i}(R) = 0$ for all $R > R^s_i$, a contradiction.[3]

Convergence by $\{\pi^s_{1it}\}$ and $\{R^s_{it}\}$ means

$$\lim_{t\to\infty}\left|\pi^s_{1it+1}(R^s_{it+1}) - \pi^s_{it}(R^s_{it})\right| = 0.$$

Yet Equation (A.7) implies

$$\begin{aligned}&\lim_{t\to\infty}\left|\pi^s_{1it+1}(R^s_{it+1}) - \pi^s_{it}(R^s_{it})\right|\\ &\quad= \lim_{\varepsilon\to 0}\frac{1}{u_i}\Big(\left[c'_i(R^s_{it}+\varepsilon) - c'_i(R^s_{it})\right] - \left[\phi'_i(R^s_{it}+\varepsilon) - \phi'_i(R^s_{it})\right]\Big)\\ &\quad= \frac{1}{u_i}\left(c''_i(R^s_i) - \phi''_i(R^s_i)\right),\end{aligned} \tag{A.8}$$

[3]The assumption that $i$ updates subjective probabilites based on knowlege acquired from his own experience in peer review implies that, if $i$ is accepted at stage 1 in time $t'$ for review group $s$, then $\Pi^s_{1it}(R) = 1$ for all $t > t'$ and $R \geq R_{it'}$; otherwise, $\Pi^s_{1it}(R) = 0$ for all $t > t'$ and $R \leq R_{it'}$. Similarly, if $i$ receives an R&R at stage 0 in time $t'$ for review group $s$, then $\Pi^s_{0it}(r) = 1$ for all $t > t'$ and $r \geq r_{0it'}$; otherwise, $\Pi^s_{it}(r) \leq \Pi^s_{it'}(r)$ for all $t > t'$, $r \leq r_{0it'}$ and $s \in \Sigma$.

where $R^s_{it} \to R^s_i$ guarantees that for all (sufficiently small) $\varepsilon > 0$ there exists $R^s_{it+1} = R^s_{it} + \varepsilon$. $u_i > 0$, $c_i''(R) > 0$ and $\phi_i''(R) < 0$ by assumption; thus, Equation (A.8) is strictly positive. According to Equation (A.8), $\{\pi^s_{1it}\}$ does not converge, a contradiction.

Consider now the proposition that $r_{0i} \leq r_{0it}$ for all $t$ past some $t''$. As before, I proceed with a contradiction. Suppose $r_{0it} < r_{0i}$ for all $t > t'$, where $t'$ is large enough that $\widetilde{r}^q_{0i} \notin (r_{0it'}, r_{0i})$ for all $q \neq s$ and $r^s_{1it+1} \leq r^s_{1it}$ for all $s \in \Sigma$.

At time $t$, the author chooses $r_{0it}$. This choice is governed by the first-order condition in Equation (A.1):

$$K + \mu^s_i \left( \pi^s_{0it}(r_{0it}) v^s_{1it} + \Pi^s_{0it}(r_{0it}) \frac{\partial v^s_{1it}}{\partial r_{0it}} \right) = c_i'(r_{0it}) - \phi_i'(r_{0it}) \tag{A.9}$$

where $\mu^s_i$ is the probability of drawing state $s$ and

$$K = \int_{\Sigma \setminus s} \left( \pi^q_{0it}(r_{0it}) v^q_{1it} + \Pi^q_{0it}(r_{0it}) \frac{\partial v^q_{1it}}{\partial r_{0it}} \right) \mathrm{d}\mu_i$$

is the marginal change in expected stage 1 subjective utility in all states $q \neq s$.

If $r^s_{1it+1} > 0$ then $r^s_{1it} > 0$. Thus $\partial v^s_{1it} / \partial r_{1it} = 0$; from Equation (A.2), Equation (A.9) is equivalent to

$$K + \mu^s_i \pi^s_{0it}(r_{0it}) v^s_{1it} = \Big( 1 - \mu^s_i \Pi^s_{0it}(r_{0it}) \Big) \Big( c_i'(r_{0it}) - \phi_i'(r_{0it}) \Big). \tag{A.10}$$

If $r^s_{1it} = 0$ then $r^s_{1it+1} = 0$, and $\partial v^s_{1it} / \partial r_{1it} = \pi^s_{1it}(R^s_{it}) u_i$.[4] In this case, Equation (A.9) is equivalent to

$$K + \mu^s_i \Big( \pi^s_{0it}(r_{0it}) v^s_{1it} + \Pi^s_{0it}(r_{0it}) \pi^s_{1it}(R^s_{it}) u_i \Big) = c_i'(r_{0it}) - \phi_i'(r_{0it}). \tag{A.11}$$

By the monotone convergence theorem, $\{v^s_{1it}\}$ and $\{\Pi^s_{0it}\}$ converge.[5] If $\widetilde{r}^s_{0i} < r_{0i}$, then $\pi^s_{0it}(r_{0it}) = 0$ for all $t > t'$, where $t'$ has been redefined to assure $\widetilde{r}^s_{0i} \leq r_{0it}$; if $r_{0i} \leq \widetilde{r}^s_{0i}$, then

$$\lim_{t \to \infty} \Pi^s_{0it}(r_{0it}) = \lim_{t \to \infty} \sum_{r \in \Omega_t} \pi^s_{0it}(r) = \pi^s_{0i}(r_{0i}), \tag{A.12}$$

where $\Omega_t = (r_{0it-1}, r_{0it}]$. $\pi^s_{0i}(r_{0i}) = \infty$ implies $\lim \Pi^s_{0it} = \infty$, which is impossible given $\Pi^s_{0it}$, by definition, is a bounded function. Hence, $\{\pi^s_{0it}\}$ is likewise convergent, so

$$\begin{aligned} &\lim_{t \to \infty} \left| \mu^s_i \left( \pi^s_{0it+1}(r_{0it+1}) v^s_{1it+1} - \pi^s_{0it}(r_{0it}) v^s_{1it} \right) \right| \\ &\quad = \mu^s_i \left( \lim_{t \to \infty} \pi^s_{0it+1}(r_{0it+1}) \lim_{t \to \infty} v^s_{1it+1} - \lim_{t \to \infty} \pi^s_{0it}(r_{0it}) \lim_{t \to \infty} v^s_{1it} \right) \\ &\quad = 0 \end{aligned}$$

and

$$\begin{aligned} &\lim_{t \to \infty} \left| \mu^s_i u_i \left( \Pi^s_{0it+1}(r_{0it+1}) \pi^s_{1it+1}(R^s_{it+1}) - \Pi^s_{0it}(r_{0it}) \pi^s_{1it}(R^s_{it}) \right) \right| \\ &\quad = \mu^s_i \, u_i \left( \lim_{t \to \infty} \Pi^s_{0it+1}(r_{0it+1}) \lim_{t \to \infty} \pi^s_{1it+1}(R^s_{it+1}) - \lim_{t \to \infty} \Pi^s_{0it}(r_{0it}) \lim_{t \to \infty} \pi^s_{1it}(R^s_{it}) \right) \\ &\quad = 0. \end{aligned}$$

[4] If $r^s_{1it} > 0$ and $r^s_{1it+1} = 0$, redefine $t'$ as $t' + 1$. $r^s_{1it+1} \leq r^s_{1it+1}$ for all $t > t'$ precludes $r^s_{1it} = 0$ and $r^s_{1it+1} > 0$.

[5] $\partial v^s_{1it} / \partial r_{0it} \geq 0$ and $v^s_{1it}$ is bounded below by zero and above by $u_i + \max\{\phi_i(R^\star_i) - c_i(R^\star_i), 0\}$. $\pi^s_{0it}(r_{0it}) \geq 0$ since $r_{0it} < r_{0it+1}$ (by assumption) and $\Pi^s_{0it}$ is bounded by 0 and 1 (by definition).

For the moment, assume there exists $t''$ such that for all $r \in (r_{0it''}, r_{0i})$, $K$ is constant.[6] Thus, changes over time to the left-hand sides of Equation (A.10) and Equation (A.11) converge to 0. Yet the right-hand sides of Equation (A.10) and Equation (A.11) do not, since

$$\lim_{t\to\infty} \mu_i^s \Pi_{0it}^s(r_{0it}) = \mu_i^s \Pi_{0i}^s(r_{0i})$$

is strictly less than 1, where $\Pi_{0i}^s$ is the finite limit of $\{\Pi_{0it}^s\}$, while

$$\begin{aligned}
&\lim_{t\to\infty} \left| \big(c_i'(r_{0it+1}) - c_i'(r_{0it})\big) - \big(\phi_i'(r_{0it+1}) - \phi_i'(r_{0it})\big) \right| \\
&\quad = \lim_{\varepsilon\to 0} \big(c_i'(r_{0it}+\varepsilon) - c_i'(r_{0it})\big) - \big(\phi_i'(r_{0it}+\varepsilon) - \phi_i'(r_{0it})\big) \\
&\quad = c_i''(r_{0i}) - \phi_i''(r_{0i})
\end{aligned}$$

is strictly greater than 0, where convergence of $\{r_{0it}\}$ guarantees that for all (sufficiently small) $\varepsilon > 0$ there exists $r_{0it+1} = r_{0it} + \varepsilon$.[7] Thus, a contradiction.

Although the contradiction depends on the existence of $t''$, the finite sum of convergent sequences is also convergent. Thus, for any finite number of states in which $\pi_{0it}^q \neq 0$ changes to the left-hand sides of Equation (A.10) and Equation (A.11) converge to 0 while changes to their right-hand sides do not. Because the number of states is finite by assumption, this establishes the general contradiction. □

**Lemma 3.** $\Pi_{0it}^s(r_{0it}) \to \mathbf{1}_{0i}^s(r_{0i})$ *and* $\Pi_{1it}^s(R_{it}^s) \to \mathbf{1}_{1i}^s(R_i^s)$.

*Proof.* As established in Lemma 2, $R_i^s \leq R_{it}^s$ for all $t > t''$. If $R_i^s < \widetilde{R}_i^s$ then $R_{it}^s < \widetilde{R}_i^s$ for all $t > t''$ where $t''$ has been redefined to satisfy the latter inequality. Thus, the paper is rejected for all $t > t''$ and $\Pi_{1it}^s(R) = 0$ for all $R \leq R_{it''}^s$ and $t > t''$. If $\widetilde{R}_i^s \leq R_i^s$, then $\widetilde{R}_i^s \leq R_{it}^s$ for all $t > t''$ (again $t''$ redefined to satisfy this inequality). Thus, the paper is accepted for all $t > t''$. $\Pi_{1it+1}^s(R) = 1$ for all $R \geq R_{it}^s$ and $t > t''$; $\Pi_{1it}^s(R_{it}^s)$ converges to 1 at the limit.

Also from Lemma 2, $r_{0i} \leq r_{0it}$ for all $t > t'$. If $r_{0i} < \widetilde{r}_{0i}^s$, then the paper is rejected at stage 0 for all $t > t'$, where $t'$ is defined so that $r_{0it} < \widetilde{r}_{0i}^s$ for all $t > t'$. Define $t'' > t'$ such that for all $t > t''$, the probability of having reached state $s$ is 1; thus, $\Pi_{it}^s(r_{0it}) = 0$ for all $t > t''$. If $\widetilde{r}_{0i}^s \leq r_{0i}$, then redefine $t''$ so that $\widetilde{r}_{0i}^s \leq r_{0it}$ for all $t > t''$. The paper is accepted, $s$ is revealed and $\Pi_{0it+1}^s(r) = 1$ for all $r \geq r_{0it}$ and $t > t''$; $\Pi_{0it}^s(r_{0i})$ converges to 1 at the limit. Thus, all is proved. □

**Lemma 4.** *There exists a unique cluster point of* $\{(r_{0it}, R_{it}^{s^\star})\}$ *for every* $s^\star \in \Sigma$.

*Proof.* Suppose $\{(r_{0it}, R_{it}^{s^\star})\}$ has two cluster points: $(r_{0i}', R_i^{s'})$ and $(r_{0i}'', R_i^{s''})$. Denote their respective convergent subsequences by $\{(r_{0it}', R_{it}^{s'})\}$ and $\{(r_{0it}'', R_{it}^{s''})\}$. Given the concavity of $\phi_i$ and convexity of $c_i$, a unique readability at each stage maximises Equation (8) and Equation (9) for fixed $\Pi_{0it}^s$ and $\Pi_{1it}^s$. Thus, $r_{0i0}' = r_{0i0}''$ and $R_{i0}^{s'} = R_{i0}^{s''}$ at time 0.

Assume at time $t$ the author has chosen $r_{0il}' = r_{0il}''$ and $R_{il}^{s'} = R_{il}^{s''}$ for all $l < t$; thus, $\Pi_{0it}^{s'}(r) = \Pi_{0it}^{s''}(r)$ and $\Pi_{1it}^{s'}(R) = \Pi_{1it}^{s''}(R)$ for all $r$ and $R$, so the author chooses $r_{0it}' = r_{0it}''$ and $R_{it}^{s'} = R_{it}^{s''}$ at time $t$ as well. By the axiom of induction, $\{(r_{0it}', R_{it}^{s'})\} = \{(r_{0it}'', R_{it}^{s''})\}$ for all $t$ so $(r_{0i}, R_i^s)$ is unique.[8] Since the choice of $s$ was arbitrary, there exists a unique cluster point of $\{(r_{0it}, R_{it}^{s^\star})\}$ for every $s^\star \in \Sigma$. □

**Lemma 5.** *Consider two equivalent authors,* $i$ *and* $k$*, such that*

[6] Effectively, this assumes $\pi_{0it}^q(r) = 0$ for all $r \in (r_{0it''}, r_{0i})$ and $q \neq s$ and (i) $\Pi_{0it}^q(r) = 0$ for all $q$ in which $r_{0i} < \widetilde{r}_{0i}^q$; (ii) $\Pi_{0it}^q(r) = 1$ and $\pi_{1it}^q(R_{it}^q) = 0$ for all $q$ in which $\widetilde{r}_{0i}^q < r_{0i}$; and (iii) $\widetilde{r}_{0i}^q \neq r_{0i}$ for any $q$. Collectively, these assumptions imply convergence of $\{\pi_{0it}^q\}$, $\{R_{it}^q\}$ and $\{\pi_{1it}^q\}$ in every state $q \neq s$ and no change to the author's marginal stage 1 objective function given a small increase in $r$ in any state but $s$.

[7] Although the change in $1 - \mu_i^s \Pi_{0it}^s(r_{0it})$ between time $t$ and $t+1$ converges to 0, it cannot converge faster than $c_i'(r_{0it}) - \phi_i'(r_{0it})$ unless $\pi_{0it}^s(r_{0i}) = \infty$, which Equation (A.12) shows is not possible.

[8] Note that $r_{0it}$ is chosen before $s$ is realised, meaning $r_{0i}$ is the unique cluster point of $\{r_{0it}\}$ regardless of $s$.

1. *for at least one* $t'' < t'$, $(r_{0it''}, R_{it''}) < (r_{0it'}, R_{it'})$ *and there exists* $K'' > 0$ *such that for no* $t > t'$, $||(r_{0it}, R_{it}) - (r_{0it''}, R_{it''})|| < K''$*; and*

2. $(r_{0kt}, R_{kt}) \leq (r_{0it}, R_{it})$ *for all* $s \in \Sigma_{A_{it}}$ *and* $t > t'$ *and there exists* $K' > 0$ *such that for at least one* $s \in \Sigma_{A_{it}}$ *and no* $t > t'$, $||(r_{0it}, R_{it}) - (r_{0kt}, R_{kt})|| < K'$.

*If* $\widetilde{r}^s_{0i} = \widetilde{r}^s_{0k}$, $\widetilde{R}^s_i = \widetilde{R}^s_k$ *and* $\mu^s_i = \mu^s_k$ *for all* $s \in \Sigma$*, then*

$$\int_\Sigma \mathbf{1}^s_{0k}(r_{0kt})\mathbf{1}^s_{1k}(R_{kt})\,\mathrm{d}\mu_k < \int_\Sigma \mathbf{1}^s_{0i}(r_{0it})\mathbf{1}^s_{1it}(R_{it})\,\mathrm{d}\mu_i. \tag{A.13}$$

*Proof.* Suppose for the moment that $\Sigma_{A_{it}}$ contains only state $q$ and assume $r_{0kt} = r_{0it}$. Since $q$ is the only state in $\Sigma_{A_{it}}$, $R^q_{kt} < R^q_{it}$. As a result,

$$\mathbf{1}^s_{0k}(r_{0kt})\mathbf{1}^s_{1k}(R^s_{kt}) = \mathbf{1}^s_{0i}(r_{0it})\mathbf{1}^s_{1i}(R^s_{it}) = 0 \text{ for all } s \neq q,$$

and

$$\mathbf{1}^s_{0k}(r_{0kt})\mathbf{1}^s_{1k}(R^s_{kt}) \leq \mathbf{1}^s_{0i}(r_{0it})\mathbf{1}^s_{1i}(R^s_{it}) = 1 \text{ for } s = q. \tag{A.14}$$

If I show that the inequality in Equation (A.14) is strict, then Equation (A.13) is true. By way of a contradiction, assume it holds as an equality. Thus, $\widetilde{R}^q_i \leq R^q_k < R^q_i$, where $R^q_{kt} \to R^q_k$ and $R^q_{it} \to R^q_i$ (Lemma 4). Together with $R^\star_i \leq r_{0it''} < R^q_i$, this implies

$$\lim_{\varepsilon \to 0-} \Pi^q_{1i}(R^q_i + \varepsilon) < 1.^{9} \tag{A.15}$$

Meanwhile, author $i$ observes author $k$'s prior readability choices, publication history and paper count.[10] From this, he discovers

$$\lim_{N_k \to \infty} \frac{N_{A_k}}{N_k} = \mu^q_i, \tag{A.16}$$

where $N_{A_k}$ and $N_k$ are author $k$'s accepted and total paper counts, respectively. Because $i$ updates $\Pi^s_{1it}$ when he observes with probability 1 that in state $s$, $k$ is accepted at some $R \neq R^s_i$, Equation (A.16) necessarily implies

$$\lim_{\varepsilon \to 0-} \Pi^s_{1i}(R^s_i + \varepsilon) = 1,$$

a contradiction.

Similar proofs by contradiction show that the inequality in Equation (A.14) must also be strict when $R^q_{kt} = R^q_{it}$ and $r_{0kt} < r_{0it}$ in state $q$ and when $\Sigma_{A_{it}}$ contains more than one state. □

*Proof of Corollary 1.* Let

$$r_{0it} = \max\left\{R^\star_i, \widetilde{r}^{\overline{s}}_{0i} + e^{\overline{s}}_{0it}\right\} \quad \text{and} \quad R_{it} = \max\left\{r_{0it}, \widetilde{R}^s_i + e^s_{1it}\right\}, \tag{A.17}$$

and define $\delta^s_{0ik}$ and $\delta^s_{1ik}$ as the difference in readability standards applied to authors $i$ and $k$ by review group $s$ in time $t$ at stage 0 and 1, respectively:

$$\delta^s_{0ik} \equiv \widetilde{r}^s_{0i} - \widetilde{r}^s_{0k} \quad \text{and} \quad \delta^s_{1ik} \equiv \widetilde{R}^s_i - \widetilde{R}^s_k.$$

[9] That is, $\Pi^q_{0i}(R) = 1$ for all $R \geq R^q_i$. Because he chose $R^\star_i \leq R_{it''} < R^q_i$ at some earlier date, the author's marginal benefit from a higher $R$ is decreasing when the probability of acceptance remains constant. Thus, if he optimally chooses $R^q_i > \max\{R_{it''}, R^q_k\}$, it must be because there is no smaller $R$ that satisfies Equation (A.7). This is only possible if there is a jump discontinuity in $\Pi^q_{0i}$ at $R^q_i$, as illustrated in Equation (A.15).

[10] The assumption that $i$ updates subjective probabilities based on conclusive evidence derived from the choices and outcomes of equivalent peers implies that, if $i$ observes with probability 1 that in state $s$ an equivalent author $k$ receives an R&R at $r_{0k}$, then $\Pi^s_{0it}(r) = 1$ for all $r \geq r_{0k}$. Similarly, if $i$ observes with probability 1 that in state $s$, $k$ is accepted at $R_k$, then $\Pi^s_{1it}(R) = 1$ for all $R \geq R_k$.

I first show that Equation (10) conservatively estimates $D_{ik}$ when $\Sigma_{A_{it}} \subset \Sigma_{A_{kt}}$. Let $r_{0it} < R_{it}$. From Equation (A.17) and the definition of $\delta^s_{1ik}$,

$$\begin{aligned} R_{it} - R_{kt} &= \widetilde{R}^s_i + e_{1it} - \max\left\{R^\star_k, \widetilde{r}^{\overline{s}_k}_{0k} + e_{0kt}, \widetilde{R}^s_k + e_{1kt}\right\} \\ &\leq \widetilde{R}^s_i - \widetilde{R}^s_k + e_{1it} - e_{1kt} \\ &= \delta^s_{1ik} + e_{1it} - e_{1kt}, \end{aligned} \tag{A.18}$$

where $\overline{s}_k$ is the review group in $\Sigma_{A_{kt}}$ for which $\widetilde{r}^s_{0k}$ is highest. When $R_{it} = r_{0it}$, however, Equation (A.17) and the definition of $\delta^s_{0ik}$ instead imply:

$$\begin{aligned} R_{it} - R_{kt} &= \max\left\{R^\star_i, \widetilde{r}^{\overline{s}_i}_{0i} + e_{0it}\right\} - \max\left\{R^\star_k, \widetilde{r}^{\overline{s}_k}_{0k} + e_{0kt}, \widetilde{R}^s_k + e_{1kt}\right\} \\ &\leq \max\left\{R^\star_i, \widetilde{r}^{\overline{s}_i}_{0i} + e_{0it}\right\} - \widetilde{r}^{\overline{s}_k}_{0k} - e_{0kt}, \end{aligned} \tag{A.19}$$

where $\overline{s}_i$ is the review group in $\Sigma_{A_{it}}$ for which $\widetilde{r}^s_{0i}$ is highest. From Theorem 1's second condition, $R_{it''} < R_{it}$ for some $t'' < t$. Thus, $R_{it''} < r_{0it}$. Because $R^\star_i$ is a lower bound on $r_{0it}$ for all $s$ and $t$ (Lemma 1), $R^\star_i < r_{0it}$; Equation (A.19) is equivalent to

$$\begin{aligned} R_{it} - R_{kt} &\leq \widetilde{r}^{\overline{s}_i}_{0i} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt} \\ &= \delta^{\overline{s}_i}_{0ik} + \widetilde{r}^{\overline{s}_i}_{0k} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt}. \end{aligned} \tag{A.20}$$

$e_{0it} = e_{0kt}$ and $e_{1it} = e_{1kt}$ (by assumption). Because $\Sigma_{A_{it}} \subset \Sigma_{A_{kt}}$, $\widetilde{r}^{\overline{s}_i}_{0k} \leq \widetilde{r}^{\overline{s}_k}_{0k}$ (by definition); Equation (A.20) implies $R_{it} - R_{kt} \leq \delta^{\overline{s}_i}_{0ik}$ if $R_{it} = r_{0it}$. Meanwhile, Equation (A.18) implies $R_{it} - R_{kt} \leq \delta^s_{1ik}$ if $r_{0it} < R_{it}$.

It remains to show that Equation (10) conservatively estimates $D_{ik}$ under Theorem 1's weaker Condition 3. Let $R_{it''} \leq R_{kt}$. Differences in $i$ and $k$'s preferences might influence readability—but only up to $R_{it''}$. $R_{it''} < R_{it}$ is motivated by $i$'s desire to increase his acceptance rate. Since $i$'s unconditional acceptance rate is identical to $k$'s, any $s'$ in $\Sigma_{A_{it}}$ but not in $\Sigma_{A_{kt}}$—*e.g.*, because $i$'s utility of acceptance is higher or cost of writing lower—is perfectly offset by some other $s''$ such that—because $s''$ discriminates against $i$—$s''$ is in $\Sigma_{A_{kt}}$ but not in $\Sigma_{A_{it}}$. Thus, $R_{it} - R_{kt}$ remains a conservative estimate $D_{ik}$.

Now let $R_{kt} < R_{it''}$. Since $i$'s unconditional acceptance rate at $R_{it}$ is identical to $k$'s at $R_{kt}$, $k$'s acceptance rate at $R_{it''}$ must be at least as high as $i$'s at $R_{it}$. Without loss of generality, assume they are identical. Preferences are time independent, so holding acceptance rates constant, $i$ prefers $R_{it''}$ to $R_{it}$. A time $t$ choice of $R_{it}$ over $R_{it''}$ reveals a higher probability of acceptance for the former—and a necessarily lower probability of acceptance for $i$ than $k$ at $R_{it''}$. Given $i$ and $k$ are equivalent, this difference is due to $\delta^{\overline{s}_i}_{0ik}$ or $\delta^s_{1ik}$. $R_{it} - R_{it''}$ is a conservative estimate of $R_{ik}$. Thus, all is proved.

□

## B Readability scores

### *B.1 Validity*

Advanced vocabulary and complicated sentences are the two strongest predictors of text difficulty (Chall and Dale, 1995). Hundreds of readability formulas exploit this relationship. In this paper, I concentrate on the five most widely used, tested and reliable formulas for adult reading material: Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG (Simple Measure of Gobbledegook) and Dale-Chall (DuBay, 2004). Each are listed in Figure B.1.

| Score | Formula |
|---|---|
| Flesch Reading Ease | $206.84 - 1.02 \times \frac{\text{words}}{\text{sentences}} - 84.60 \times \frac{\text{syllables}}{\text{words}}$ |
| Flesch-Kincaid | $-15.59 + 0.39 \times \frac{\text{words}}{\text{sentences}} + 11.80 \times \frac{\text{syllables}}{\text{words}}$ |
| Gunning Fog | $0.40 \times \left(\frac{\text{words}}{\text{sentences}} + 100 \times \frac{\text{polysyllabic words}}{\text{words}}\right)$ |
| SMOG | $3.13 + 5.71 \times \sqrt{\frac{\text{polysyllabic words}}{\text{sentences}}}$ |
| Dale-Chall | $3.64 + 0.05 \times \frac{\text{words}}{\text{sentences}} + 15.79 \times \frac{\text{difficult words}}{\text{words}}$ |

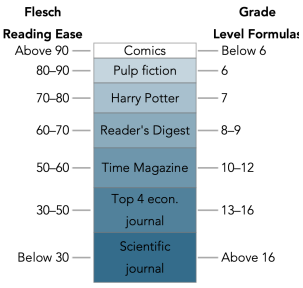


FIGURE B.1: Calculating and interpreting readability scores

*Notes.* Left-hand table displays formulas used to calculate readability scores. Polysyllabic words refer to words with three or more syllables; difficult words are those not found on a list of 3,000 words understood by 80 percent of fourth-grade readers (aged 9–10) (Chall and Dale, 1995). The graphic on the right provides a rough guide for interpreting the scores (adapted from Flesch, 1949).

The Flesch Reading Ease formula ranks passages of text in ascending order—*i.e.*, more readable passages earn higher scores. The other four formulas generate grade levels estimating the minimum years of schooling necessary to confidently understand an evaluated text—and so more readable passages earn lower scores. To minimise confusion, I multiply the four grade-level scores by negative one. Thus, higher numbers universally correspond to clearer writing throughout this paper.

The constants in each formula vary widely as do the components used to rank vocabulary. Because of these differences, grade-level scores rarely generate identical figures. Nevertheless, all five scores produce similar rankings. The yellow box plot in Figure B.2 summarises 169 inter-score correlations found in 26 studies. The median is 0.87.

Readability scores correlate with (i) oral reading fluency, (ii) human judgement, (iii) reading comprehension tests and (iv) the cloze procedure.[11] The dark blue box plots in Figure B.2 summarise 167 correlations in 38 published cross-validation studies.

Other studies have validated readability scores against surrogate measures of reading comprehension. More readable high school and college-level correspondence courses have higher completion rates (Klare and Smart, 1973). More readable academic journals enjoy larger readerships (Richardson, 1977; Swanson, 1948); their most readable articles win more awards (Sawyer et al., 2008) and are downloaded more often (Guerini et al., 2012).

More readable abstracts are also (generally) cited more frequently (see Dowling et al. (2018) and McCannon (2019) and Figure B.2). They are also more likely to be published in top-five and other higher ranking journals (Marino Fages, 2020). In a blog post, Lukas Püttmann compares abstract readability to page views of VoxEU.org columns: more readable columns are viewed three percent more often (Püttmann, 2017). Evidence from other studies linking readability and citations is, however, weaker (Berninger et al., 2017; Laband and Taylor, 1992; Lei and Yan, 2016). My own data suggest a positive relationship in papers published after 1990—and particularly those published post–2000—but no relationship before that (Figure B.2).

[11] Oral reading fluency is generally measured as the number of words read aloud correctly per minute. The cloze procedure ranks passages of text according to average readers' ability to correctly guess randomly deleted words.

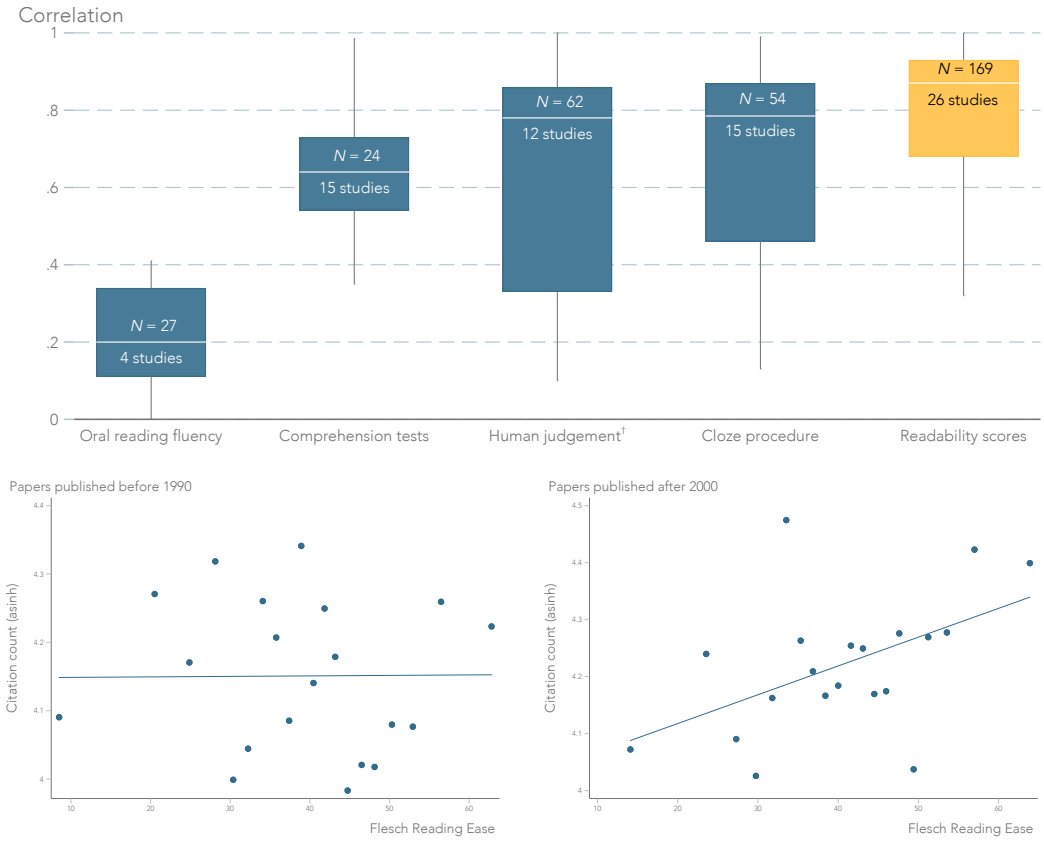


FIGURE B.2: Readability score validity

*Notes*. Top figure displays box plots of correlations between alternative measures of text difficulty and the Flesch Reading Ease, Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall readability scores. It includes 336 correlations found in 55 mostly peer reviewed papers. (See Appendix B.5 for the list of included studies and information on how they were selected.) Bottom figures plot abstracts' Flesch Reading Ease scores against their articles' citation counts (inverse hyperbolic sine (asinh) transformation) for the samples of top-four (excluding *AER Papers & Proceedings*) articles published before 1990 (left; 3,732 articles) and post-2000 (right; 3,410 articles). Each point represents the mean (in both dimensions) of roughly 170–180 observations. †Includes two studies which assessed readability using the Readability Assessment INstrument (RAIN), a comprehensive framework based on 14 variables, *e.g.*, coeherence, writing style, illustrations and typography.

Thanks to high predictive power and ease of use, readability formulas are widely employed in education, business and government. The U.S. Securities and Exchange Commission encourages clearer financial disclosure forms benchmarked against the Gunning Fog, Flesch-Kincaid and Flesch Reading Ease scores (Cox, 2007). The formulas have also guided readability assessments of, *inter alia*, standardised test questions (Chall et al., 1983; Chall et al., 1977), medical inserts (*e.g.*, Wallace et al., 2008), technical manuals (*e.g.*, Hussin et al., 2012; Klare and Smart, 1973), health pamphlets (*e.g.*, Foster and Rhoney, 2002; Meade and Byrd, 1989) and data security policies (Alkhurayyif and Weir, 2017).

In research, readability scores are considered objective proxies for "complexity". Enke (2018) controls for language sophistication using the Flesch Reading Ease formula in a study of moral values in U.S. presidential elections. Spirling (2016) employs the same score to show that British parliamentarians simplified speeches to appeal to less educated voters in the wake of the Great Reform Act. Legal research has found that judges are more reliant on legislative history when interpreting complex legal statutes, as measured by the Flesch-Kincaid formula (Law and Zaring, 2010). In finance, the scores have linked clarity of financial communication to better firm and market financial health (Biddle et al., 2009; Jansen, 2011; Li, 2008), larger investment and trading volume (De Franco et al., 2015; Lawrence, 2013; Miller, 2010; Thörnqvist, 2015) and lower demand for—albeit higher reliability of—outside research by sell-side analysts (Lehavy et al., 2011).[12]

[12]See also Loughran and McDonald (2016) for a review of the use of readability scores in finance and accounting research.

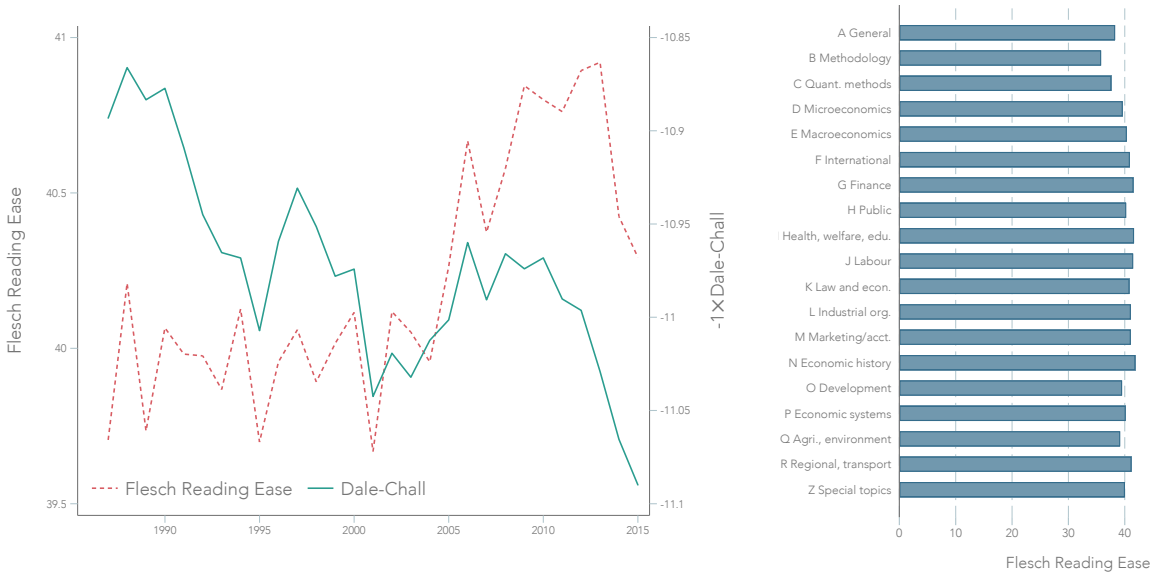


FIGURE B.3: Readability by year and *JEL* code

*Notes.* Figure on the left displays five-year moving averages of abstracts' Flesch Reading Ease (left axis) and $-1 \times$ Dale-Chall (right axis) readability scores. Sample restricted to the years 1987–2015 (6,176 articles). Figure on the right displays abstracts' Flesch Reading Ease scores averaged over primary *JEL* classifications. Data only available after 1990 (5,216 articles).

### *B.2 Descriptive statistics*

Figure B.3 breaks down the sample's abstract readability by publication year and primary *JEL* classification. Table B.1 displays data coverage by journal and decade.

The left-hand graph in Figure B.3 displays readability scores (Flesch Reading Ease and $-1 \times$ Dale Chall) between 1987–2015. Neither scores have changed drastically over the roughly 30-year period. The right-hand graph in Figure B.3 shows the Flesch Reading Ease broken down by primary *JEL* classification. There is some slight evidence that papers in Economic History are better written whereas those in quantitative methods less so. Otherwise, the score does not appear to differ much by field.

TABLE B.1: Article count, by journal and decade

| Decade | *AER* | *ECA* | *JPE* | *QJE* | Total |
|---|---|---|---|---|---|
| 1950–59 | | 120 | | | 120 |
| 1960–69 | | 343 | 184 | | 527 |
| 1970–79 | | 660 | 633 | 1 | 1,294 |
| 1980–89 | 180 | 648 | 562 | 401 | 1,791 |
| 1990–99 | 476 | 443 | 478 | 409 | 1,806 |
| 2000–09 | 693 | 520 | 408 | 413 | 2,034 |
| 2010–15 | 732 | 384 | 181 | 251 | 1,548 |
| Total | 2,081 | 3,118 | 2,446 | 1,475 | 9,120 |

*Notes.* Included is every article published between January 1950 and December 2015 for which an English abstract was found (i) on journal websites or websites of third party digital libraries or (ii) printed in the article itself. Papers published in the May issue of *AER* (*Papers & Proceedings*) are excluded. Final row and column display total article counts by journal and decade, respectively.

### *B.3 Measurement error*

Readability scores fail to capture many elements relevant to reading comprehension, including grammar—*e.g.*, active vs. passive tense (Coleman, 1964; Coleman, 1965)—legibility—*e.g.*, typeface or

layout—and content—*e.g.*, coherence, organisation and general appeal (Armbruster, 1984; Kemper, 1983; Kintsch and Miller, 1984; Meyer, 1982). Nevertheless, "long sentences generally correspond to complex syntactic structures, infrequent words generally refer to complex concepts, and hard texts will generally lead to harder questions about their content" (Kintsch and Miller, 1984, p. 222).

Still, readability scores' low causal power raises legitimate concerns about measurement error. As long as this error does not partially correlate with the variable of interest (gender), the analytical results I present in this paper attenuate toward zero (classical measurement error). Unfortunately, they are systematically biased in an unknown direction if it does (non-classical measurement error).

Sources of non-classical measurement error are threefold: (a) grammatical, spelling and transcription errors in the textual input; (b) errors in the estimates of vocabulary complexity and sentence length introduced by automating their calculation; or (c) embodied in the jump from using these two variables to infer readability.

Conditional on accurate calculation, readability scores combine very precise estimates of vocabulary complexity with almost perfect measures of sentence length (for a discussion, see Chall and Dale, 1995). The weighted average of these two variables is informative in much the same way that inferences about readability are. Thus, measurement error related to (c) should only shift superficial interpretation of observed gender differences—from "women are better writers" to "women use simpler words and write shorter sentences"—but leave conclusions deduced from them intact.

Nevertheless, I try to minimise measurement error from (c) by using abstracts as textual input. Abstracts are self-contained, universally summarise the research and are the first and most frequently read part of an article (King et al., 2006). Additionally, they follow a more standardised layout compared to other parts of a manuscript: they are generally surrounded by ample whitespace and most editorial management systems anyway reproduce them in pre-formatted cover pages. These factors suggest a relatively homogenous degree of review across journals and subject matter and limit the impact that physical layout, figures and surrounding text have on readability.

Moreover, prior research suggests authors write in a stylistically consistent manner across the abstract, introduction and discussion sections of a paper. According to an analysis of published education and psychology articles, within-manuscript correlations of Flesch Reading Ease scores range from 0.64 (abstracts vs. introductions) to 0.74 (abstracts vs. discussions) (Hartley et al., 2003). Plavén-Sigray et al. (2017) also found a strong positive correlation using full text articles from several scientific journals. Figure B.4 plots abstract readability against the readability of a passage from the introduction for 339 NBER Working Papers eventually published in a top-four journal. It suggests a similarly positive relationship holds in economics, as well.[13]

In my opinion, non-classical measurement error from (a) and (b) poses a bigger concern to the identification mapped out in this paper. I have taken several steps to reduce it. First, abstract text is also ideal for calculating readability: 100–200 words containing few score-distorting features of academic writing—*e.g.*, citations, abbreviations and equations (Dale and Chall, 1948). Additionally, most abstracts have been previously converted to accurate machine-readable text by digital libraries and bibliographic databases, curbing errors in transcription.

Second, I carefully proofread the text in order to identify (and fix) remaining transcription errors,[14] eliminate non-sentence-ending full stops, and replace typesetting code—typically used to render equations—with equivalent unicode characters.[15] Readability scores were determined using the modified text.

[13] For comparison, I randomly assigned abstracts to introductions in 1,000 simulated samples. The average coefficient of correlation between abstract text readability and the readability of a passage of text from a randomly selected introduction was –0.0006 for the Gunning Fog score and 0.0007 for the Flesch-Kincaid score.

[14] *E.g.*, words in transcribed text are often inappropriately hyphenated—typically because the word was divided at the end of the line in the original text.

[15] When no exact replacement existed, characters were chosen that mimicked as much as possible the equation's original intent while maintaining the same character and word counts. (Equations in abstracts generally only occur in *Econometrica* articles published before 1980.)

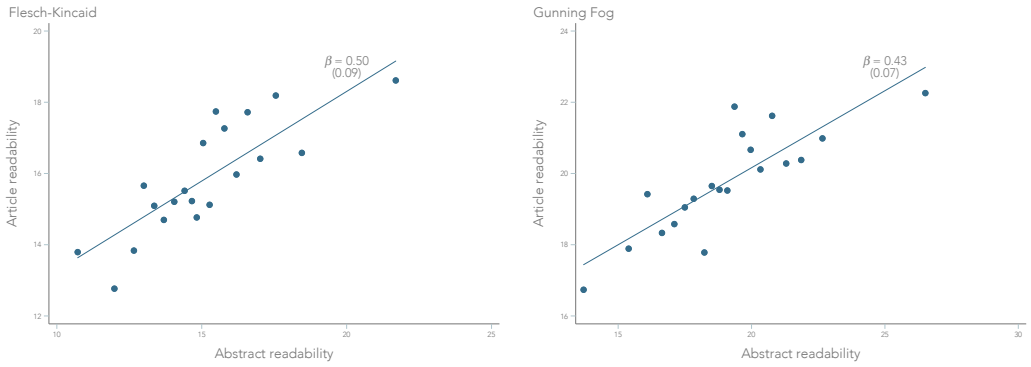


FIGURE B.4: Abstract vs. article readability

*Notes*. Figures plot abstract readability against the readability of a 150–200 word passage of text from the introduction of the same paper. $\beta$ is the slope of the regression line (robust standard errors in parentheses). Sample only includes NBER Working Papers eventually published in a top-four economics journal with a heading explicitly titled "Introduction" (339 abstract-article pairs). Data are grouped into roughly 20 equal-sized bins; each point represents the mean (in both dimensions) of about 16–17 observations. Non-abstract text kindly provided by Henrik Kleven and Dana Scott (Kleven, 2018). Readability scores calculated using the `R readability` package.

Finally, some programs that calculate scores rely on unclear, inconsistent and possibly inaccurate algorithms to count words and syllables, identify sentence terminations and check whether a word is on Dale-Chall's easy word list (for a discussion, see Sirico, 2007). To transparently handle these issues and eliminate ambiguity in how the scores were calculated, I wrote the Python module `Textatistic`. Its code and documentation are available on GitHub; a brief description is provided in Appendix B.4.

For added robustness, I also re-calculate scores and replicate most results using the `R readability` package (Appendix I). Coefficients are very similar to—and (to my chagrin) standard errors universally smaller than—those presented in the body of the paper.

### *B.4 `Textatistic`*

I wrote the Python module `Textatistic` to transparently calculate the readability scores in this study. The code and documentation are available on GitHub; I provide a brief description here.

To determine sentence count, the program replaces common abbreviations with their full text,[16] decimals with a zero and deletes question and exclamation marks used in an obvious, mid-sentence rhetorical manner.[17] The remaining full stops, exclamation and question marks are assumed to end a sentence and counted.

Next, hyphens are deleted from commonly hyphenated single words such as "co-author" and the rest are replaced with spaces, remaining punctuation is removed and words are split into an array based on whitespace. Word count is the length of that array.[18]

An attempt is made to match each word to one on an expanded Dale-Chall list. The count of difficult words is the number that are not found. This expanded list, available on GitHub, consists of 8,490 words. It is based on the original 3,000 words, but also includes verb tenses, comparative and superlative adjective forms, plural nouns, *etc.* It was created by first adding to the Dale-Chall list every conceivable alternate form of each word using Python's Pattern library. To eliminate nonsense words, the text of 94 English novels published online with Project Gutenberg were matched with words on the expanded list. Words not found in any of the novels were deleted.

Syllable counts are based on the C library `libhyphen`, an implementation of the hyphenation algorithm from Liang (1983). Liang (1983)'s algorithm is used by TEX's typesetting system. `libhyphen` is employed by most open source text processing software, including OpenOffice.

[16] Abbreviations which do not include full-stops are not altered. I manually replaced common abbreviations, such as "*i.e.*" and "U.S." with their abbreviated versions, sans full stops.

[17] For example, "?)." is replaced with ").".

[18] Per Chall and Dale (1995), hyphenated words count as two (or more) words.

### *B.5 Studies included in meta analysis*

Below are the studies included in the analysis from Figure B.2, which summarises correlations between readability scores and alternative measures of reading comprehension found in other research. A few notes on the criteria for inclusion and how some correlations were determined:

- I include only documents produced for the U.S. government or published peer reviewed studies—with the exception of the present paper, Benoit et al. (2017) and results from dissertations that were presented and discussed in a peer reviewed manuscript.
- I include a small number of studies with correlations between alternative readability measures and the number of words not listed on the Dale-Chall word list. In all other cases, however, correlations with only parts of a score (*e.g.*, syllables per words) are omitted.
- A few earlier studies calculated and listed various readability measures for many passages of text, but did not report coefficients of correlation between them. I manually calculated these correlations myself.


Ardoin, S. P. et al. (2005). "Accuracy of Readability Estimates' Predictions of CBM Performance." *School Psychology Quarterly* 20 (1), pp. 1–22.

Benoit, K., K. Munger, and A. Spirling (2017). "Measuring and Explaining Political Sophistication through Textual Complexity". Mimeo (cit. on p. 12).

Bormuth, J. R. (1966). "Readability : A New Approach". *Reading Research Quarterly* 1 (3), pp. 79–132.

Brown, J. D. (1998). "An EFL Readability Index". *JALT Journal* 20 (2), pp. 7–36.

Carver, R. P. (1974). *Improving Reading Comprehension*. Tech. rep. Washington, D.C.: American Institutes for Research in the Behavioral Sciences.

Caylor, J. S. et al. (1973). *Methodologies for Determining Reading Requirements of Military Occupational Specialties*. Tech. rep. Alexandria, Virginia: Human Resources Research Organization.

Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: Brookline Books (cit. on pp. 7, 10, 11).

Clauson, K. A., Q. Zeng-Treitler, and S. Kandula (2010). "Readability of Patient and Health Care Professional Targeted Dietary Supplement Leaflets Used for Diabetes and Chronic Fatigue Syndrome". *Journal of Alternative and Complementary Medicine* 16 (1), pp. 119–124.

Compton, D. L., A. C. Appleton, and M. K. Hosp (2004). "Exploring the Relationship Between Text-Leveling Systems and Reading Accuracy and Fluency in Second-Grade Students Who Are Average and Poor Decoders". *Learning Disabilities Research and Practice* 19 (3), pp. 176–184.

Crossley, S. A. et al. (2017). "Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas". *Discourse Processes* 54 (5-6), pp. 340–359.

Crossley, S. A., J. Greenfield, and D. S. McNamara (2008). "Assessing Text Readability Using Cognitively Based Indices". *TESOL Quarterly* 42 (3), pp. 475–493.

Cunningham, J. W., E. H. Hiebert, and H. A. Mesmer (2018). "Investigating the Validity of Two Widely Used Quantitative Text Tools". *Reading and Writing* 31 (4), pp. 813–833.

Dale, E. and J. S. Chall (1948). "A Formula for Predicting Readability". *Educational Research Bulletin* 27 (1), pp. 11–20 (cit. on p. 10).

Dale, E. and R. W. Tyler (1934). "A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability". *Library Quarterly: Information, Community, Policy* 4 (3), pp. 384–412.

Entin, E. B. and G. R. Klare (1978). "Some Inter-relationships of Readability, Cloze and Multiple Choice Scores on a Reading Comprehension Test". *Journal of Literacy Research* 10 (4), pp. 417–436.

Flesch, R. (1948). "A New Readability Yardstick". *Journal of Applied Psychology* 32 (3), pp. 221–233.

Froese, V. (1971). *Cloze Readability versus the Dale-Chall Formula*. Tech. rep. Winnipeg, Manitoba: University of Manitoba.
Fulcher, G. (1997). "Text Difficulty and Accessibility: Reading Formulae and Expert Judgement". *System* 25 (4), pp. 497–513.
Gray, W. W. and B. E. Leary (1935). *What Makes a Book Readable*. Chicago, Illinois: University of Chicago Press.
Greenfield, J. (1999). "Classic Readability Formulas in an EFL Context: Are They Valid for Japanese Speakers?" PhD thesis. Temple University.
— (2004). "Readability Formulas for EFL". *JALT Journal* 26 (1), pp. 5–24.
Guthrie, J. T. (1972). "Learnability versus Readability of Texts". *Journal of Educational Research* 65 (6), pp. 273–280.
Harris, A. J. and M. D. Jacobson (1976). "Predicting Twelfth Graders' Comprehension Scores". *Journal of Reading* 20 (1), pp. 43–46.
Harwell, M. R. et al. (1996). "Evaluating Statistics Texts Used in Education". *Journal of Educational and Behavioral Statistics* 21 (1), pp. 3–34.
Hayes, D. P., L. T. Wolfer, and M. F. Wolfe (1996). "Schoolbook Simplification and Its Relation to the Decline in SAT-Verbal Scores". *American Educational Research Journal* 33 (2), pp. 489–508.
Hengel, E. (2017). "Publishing while Female: Are Women Held to Higher Standards? Evidence From Peer Review." Cambridge Working Paper Economics: 1753.
Hull, L. C. (1979). "Beyond Readability: Measuring the Difficulty of Technical Writing". PhD thesis. Renesselaer Polytechnic Institute.
Janan, D. and D. Wray (2014). "Reassessing the Accuracy and Use of Readability Formulae". *Malaysian Journal of Learning and Instruction* 11 (1), pp. 127–145.
Jongsma, E. A. (1972). "The Difficulty of Children's Books: Librarians' Judgments vs. Formula Estimates". *Elementary English* 49 (1), pp. 20–26.
Kanouse, D. E. et al. (1981). *Informing Patients about Drugs: Summary Report on Alternative Designs for Prescription Drug Leaflets*. Tech. rep. Santa Monica, California: Rand Corporation.
Kemper, S. (1983). "Measuring the Inference Load of a Text". *Journal of Educational Psychology* 75 (3), pp. 391–401 (cit. on p. 10).
Kincaid, J. P. et al. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Tech. rep. Memphis, Tennessee: Naval Technical Training Command.
King, M. M., A. S. W. Winton, and A. D. Adkins (2003). "Assessing the Readability of Mental Health Internet Brochures for Children and Adolescents". *Journal of Child and Family Studies* 12 (1), pp. 91–99.
Klare, G. R. (1952). "Measures of the Readability of Written Communication: An Evaluation". *Journal of Educational Psychology* 43 (7), pp. 385–399.
Klingbeil, C., M. W. Speece, and H. Schubiner (1995). "Readability of Pediatric Patient Education Materials. Current Perspectives on an Old Problem." *Clinical Pediatrics* 34 (2), pp. 96–102.
Lee, W. D. and B. R. Belden (1966). "A Cross-Validation Readability Study of General Psychology Textbook Material and the Dale-Chall Readability Formula". *Journal of Educational Research* 59 (8), pp. 369–373.
Lenzner, T. (2014). "Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?" *Sociological Methods and Research* 43 (4), pp. 677–698.
Ley, P. and T. Florio (1996). "The Use of Readability formulas in Health Care". *Psychology, Health and Medicine* 1 (1), pp. 7–28.
Lorge, I. (1948). "The Lorge and Flesch Readability Formulas: A Correction". *School & Society* 67, pp. 141–142.
McLaughlin, G. (1969). "SMOG Grading: A New Readability Formula". *Journal of Reading* 12 (8), pp. 639–646.

Meade, C. D. and J. C. Byrd (1989). "Patient Literacy and the Readability of Smoking Education Literature". *American Journal of Public Health* 79 (2), pp. 204–206 (cit. on p. 8).
Meade, C. D. and C. F. Smith (1991). "Readability Formulas: Cautions and Criteria". *Patient Education and Counseling* 17 (2), pp. 153–158.
Miller, L. R. (1974). "Predictive Powers of the Flesch and Bormuth Readability Formulas". *International Journal of Business Communication* 11 (2), pp. 21–30.
Morris, L. A., A. Myers, and D. G. Thilman (1980). "Application of the Readability Concept to Patient-Oriented Drug Information". *American Journal of Health-System Pharmacy* 37 (11), pp. 1504–1509.
Powell-Smith, K. A. and K. L. Bradley-Klug (2001). "Another Look at the 'C' in CBM: Does It Really Matter if Curriculum-based Measurement Reading Probes Are Curriculum-based?" *Psychology in the Schools* 38 (4), pp. 299–312.
Powers, R. D., W. A. Sumner, and B. E. Kearl (1958). "A Recalculation of Four Readability Formulas". *Journal of Educational Psychology* 49 (2), pp. 99–105.
Russell, D. H. and H. R. Fea (1951). "Validity of Six Readability Formulas as Measures of Juvenile Fiction". *Elementary School Journal* 52 (3), pp. 136–144.
Singer, H. (1975). "The Seer Technique: A Non-Computational Procedure for Quickly Estimating Readability Level". *Journal of Reading Behavior* 7 (3), pp. 255–267.
Singh, J. (2003). "Reading Grade Level and Readability of Printed Cancer Education Materials". *Oncology Nursing Forum* 30 (5), pp. 867–870.
Štajner, S. et al. (2012). "What Can Readability Measures Really Tell Us About Text Complexity?" In: *Workshop on Natural Language Processing for Improving Textual Accessibility*, pp. 14–21.
Sullivan, R. J. (1976). "A Comparison of Results Obtained Using the Cloze Procedure with Readability Levels Using the Dale-Chall Formula on Selected University Textbooks". In: *26th Annual Meeting of the National Reading Conference*. Atlanta, Georgia.
Van Oosten, P., D. Tanghe, and V. Hoste (2010). "Towards an Improved Methodology for Automated Readability Prediction". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 775–782.
Wells, J. A. (1994). "Readability of HIV/AIDS Educational Materials: The Role of the Medium of Communication, Target Audience, and Producer Characteristics". *Patient Education and Counseling* 24 (3), pp. 249–259.
Woods, B., G. Moscardo, and T. Greenwood (1998). "A Critical Review of Readability and Comprehensibility Tests". *Journal of Tourism Studies* 9 (2), pp. 49–61.
Zheng, J. and H. Yu (2017). "Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study". *Journal of Medical Internet Research* 19 (3), pp. 1–15.

## C Description of control variables

*Institutions.* For every article I recorded authors' institutional affiliations. Individual universities in U.S. State University Systems were coded separately (*e.g.*, UCLA and UC Berkeley) but think tanks and research organisations operating under the umbrella of a single university were grouped together with that university (*e.g.*, the Cowles Foundation and Yale University). Institutions linked to multiple universities are coded as separate entities (*e.g.*, École des hautes études en sciences sociales).

In total, 1,039 different institutions were identified. For each institution, I count the number of articles in which it was listed as an affiliation in a given year and smooth the average over a five-year period.[19] Institutions are ranked on an annual basis using this figure and then grouped to create fifteen dynamic dummy variables. Institutions ranked in positions 1–9 are assigned individual dummy variables. Those in positions 10–59 are grouped in bins of 10 to form six dummy variables. Institutions ranked 60 or above were collectively grouped to form a final dummy variable.[20] When multiple institutions are associated with an observation, only the dummy variable with the highest rank is used, *i.e.*, the highest-ranked institution per author when data is analysed at the author-level and the highest-ranked institution for all authors when data is analysed at the article-level.

*Citations.* I use article citations from Web of Science. Unless otherwise mentioned, citation counts are transformed using the inverse hyperbolic sine function (asinh).

*Author prominence.* I generate 37 dummy variables that group authors by their career-total top-five journal (*AER*, *Econometrica*, *JPE*, *QJE* and *REStud*) publications as of December 2015 (denoted by max. $T_5$). For example, Jean Tirole forms one group (59 articles); James Heckman and Gene Grossman form another (34 articles).[21]

*Author seniority.* To account for author seniority, I control for an author's number of top-five (*AER*, *Econometrica*, *JPE*, *QJE* and *REStud*) publications at the time a paper was published (denoted by max. $t_5$). For co-authored articles, only the data corresponding to the most prolific author is used.[22]

*English fluency.* To account for English fluency, most regressions include a dummy variable equal to one if an article is co-authored by at least one native (or almost native) English speaker. I assume an author is "native" if he: (i) was raised in an English-speaking country; (ii) obtained all post-secondary eduction from English speaking institutions;[23] or (iii) spoke with no discernible (to me) non-native accent. This information was almost always found—by me or a research assistant—in authors' CVs, websites, Wikipedia articles, faculty bios or obituaries. In the few instances where the criteria were ambiguously satisfied—or no information was available—I asked friends and colleagues of the author or inferred English fluency from the author's first name, country of residence or surname (in that order).[24]

[19] Blank (1991) ranks institutions by National Academy of Science departmental rankings. Those and similar official rankings are based largely on the number of papers published in the journals analysed here.

[20] In a December 2017 versions of this paper (see my website), I construct a more comprehensive—but static—set of institutional controls. Results are very similar to those presented here. (See also Hengel (2016).)

[21] This quality/productivity control has several limitations: (i) it relies on publication counts—not necessarily an accurate measure of "quality"; (ii) it discounts current junior economists' productivity; and (iii) it generates somewhat inconsistent groupings—for example, two authors have published 34 articles, but only one author has published 29 (Joseph Stiglitz).

[22] In Hengel (2016, p. 42 and p. 44), I experiment with another measure of quality—the order an article appeared in an issue. It has no noticeable impact on the coefficient of interest or its standard error.

[23] Non-native speakers who meet this criteria have been continuously exposed to spoken and written English since age 18. This continuous exposure likely means they write as well as native English speakers. To qualify as an English-speaking institution, all courses—not just the course studied by an author—must be primarily taught in English. *E.g.*, McGill University is classified as English-speaking; University of Bonn is not (although most of its graduate economics instruction is in English).

[24] I also conducted a primitive surname analysis (see Hengel, 2016, pp. 35–36). It suggests that the female authors in my data are no more or less likely to be native English speakers.

*Field.* I create dummy variables corresponding to the 20 primary and over 700 tertiary *JEL* categories to control for subject matter. The *JEL* system was significantly revised in 1990; because exact mapping from one system to another is not possible, I collected these data only for articles published post-reform—about 60 percent of the dataset. Codes were recorded whenever found in the text of an article or on the websites where bibliographic information was scraped. Remaining articles were classified using codes from the American Economic Association's Econlit database.

*Editorial policy.* To control for editorial policy, I recorded editor/editorial board member names from issue mastheads. *AER* and *Econometrica* employ an individual to oversee policy. *JPE* and *QJE* do not generally name one lead editor and instead rely on boards composed of four to five faculty members at the University of Chicago and Harvard, respectively.[25] *REStud* is also headed by an editorial board, the size of which has been gradually increasing—from two members in the 1970s to 7–8 members today. Members are also located all over the world.

Editor controls are based on distinct lead editor/editorial boards—*i.e.*, they differ by at least one member. Among top four journals, 74 groups are formed in this manner. *REStud* adds another 34.[26]

*Family commitments.* To control for motherhood's impact on revision times, I recorded children's birth years for women with at least one entirely female-authored paper in *Econometrica*. I personally (and, I apologise, rather unsettlingly) gleaned this information from published profiles, CVs, acknowledgements, Wikipedia, personal websites, Facebook pages, background checks and local school district/popular extra-curricular activity websites.[27] Exact years were recorded whenever found; otherwise, they were approximated by subtracting a child's actual or estimated age from the date the source material was posted online. In several instances, I obtained this information from acquaintances, friends and colleagues or by asking the woman directly.

If an exhaustive search turned up no reference to children, I assumed the woman in question did not have any.[28]

[25] In recent years, *JPE* has been published under the aegis of a lead editor.

[26] Given the size of *Restud*'s editorial board and the fact that members serve fixed 3–4 full-year terms, editorial controls are highly correlated with year fixed effects. Moreover, unlike at *JPE* and *QJE*, editors are not located at the same institution. Thus, editor fixed effects may be less informative about editorial policy at *REStud* than they are for the other four journals.

[27] While the information I found was publicly available, I apologise for the obvious intrusion.

[28] Given its sensitive nature, children's birth years are not currently available on my website (unlike most of the other data in this paper).

## D Section 3.1, suplemental output

### *D.1 Readability differences across journals*

Table D.1 shows the coefficients on the journal dummies in column (2), Table 2. They compare *AER*'s readability to the readability of *Econometrica*, *JPE* and *QJE*.

Table D.1: Journal readability, comparisons to *AER*

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| *Econometrica* | −12.40*** | −4.42*** | −4.25*** | −2.62*** | −0.67*** |
| | (1.91) | (0.41) | (0.47) | (0.38) | (0.16) |
| *JPE* | −5.62*** | −3.99*** | −3.41*** | −1.83*** | 0.18 |
| | (1.91) | (0.41) | (0.47) | (0.38) | (0.16) |
| *QJE* | 1.54** | −0.02 | 0.30*** | 0.21*** | 0.27*** |
| | (0.60) | (0.13) | (0.09) | (0.06) | (0.05) |

*Notes*. Figures are the estimated coefficients on the journal dummy variables from (2) in Table 2. Each contrasts the readability of the journals in the left-hand column with the readability of *AER*. Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *D.2 Gender and readability, by* JEL *code*

Figure D.1 displays results from an ordinary least squares regression on the Dale-Chall score; regressors are: (i) ratio of female co-authors; (ii) dummies for each primary *JEL* code; (iii) interactions from (i) and (ii); (iv) controls for editor, journal, year, institution and English fluency; and (v) quality controls—citation count, max. $T_5$ fixed effects (author prominence) and max. $t_5$ (author seniority).[29] Due to small samples—particularly of female authors—Figure D.1 includes 561 articles from *AER Papers & Proceedings*.[30]

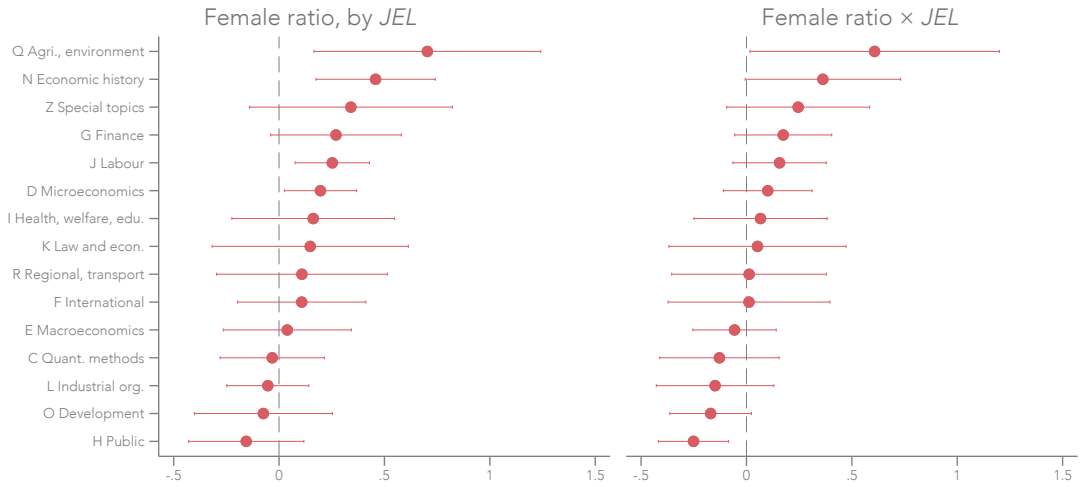


FIGURE D.1: Gender differences in readability, by *JEL* classification

*Notes.* Sample 5,777 articles, including 561 from *AER Papers & Proceedings* (see Footnote 12). Codes A, B, M and P dropped due to small sample sizes of female-authored papers (see Footnote 29). Estimates from an OLS regression of:

$$R_j = \beta_0 + \beta_1 \text{female ratio}_j + \boldsymbol{\beta}_2 \mathbf{J}_j + \boldsymbol{\beta}_3 \text{ female ratio}_j \times \mathbf{J}_j + \boldsymbol{\theta}\, \mathbf{X}_j + \varepsilon_j,$$

where $R_j$ is the readability score for article $j$; female ratio$_j$ is paper $j$'s ratio of female authors to total authors; $\mathbf{J}_j$ is a $15 \times 1$ column vector with $k$th entry a binary variable equal to one if article $j$ is classified as the $k$th *JEL* code; $\mathbf{X}_j$ is a vector of editor, journal, year, institution and English language dummies, $N_j$ (number of co-authors on paper $j$) and quality controls (citation count (asinh), max. $T_5$ fixed effects (author prominence) and max. $t_5$ (author seniority)); $\varepsilon_j$ is the error term. Left-hand graph shows marginal effects of female ratio for each *JEL* code ($\beta_1 + \beta_3^k$). The mean effect at observed *JEL* codes is 0.15 (standard error 0.049). Right-hand graph displays interaction terms ($\beta_3^k$). Horizontal lines represent 90 percent confidence intervals from standard errors adjusted for clustering on editor.

Points reflect marginal effects across *JEL* classification; bars represent 90 percent confidence intervals from standard errors clustered by editor. The mean effect at observed *JEL* codes is 0.15 (standard error 0.049). This estimate coincides with results in Table H.1—women's papers require six fewer weeks of schooling to understand—and is highly significant.

Women earn higher marks for clarity in 11 out of 15 categories; only four are at least weakly significant: Q (Agricultural and Natural Resource Economics; Environmental and Ecological Economics), N (Economic History), J (Labour Economics) and D (Microeconomics). Men may be better writers in C (Mathematical and Quantitative Methods), L (Industrial Organisation), O (Economic Development, Innovation, Technological Change, and Growth) and H (Public Economics); none, however, are statistically different from zero. Figure D.1's right-hand graph displays coefficients from interacting the ratio of female co-authors with each *JEL* code. Q and N are (weakly) significantly above the mean, H significantly below it. Remaining categories are not statistically different from the mean effect.

In general, sample sizes are small and estimates imprecise—only Labour Economics and Microeconomics contain more than 100 papers written only by women (the others average 35). Nevertheless, Figure D.1 suggests two things. First, the mostly insignificant interaction terms indicate outlier fields are probably not driving journals' gender readability gap—nor is any specific field bucking the trend.

[29] Codes A, B, M and P are dropped due to insufficient number of female-authored papers: each had fewer than 10 papers authored only by women. No paper is classified under category Y.

[30] See Hengel (2016, pp. 42–43) for a version of Figure D.1 excluding *AER Papers & Proceedings* articles.

Second, the number of women in a field appears to have little effect on the size of the gap: Agriculture/Environment has one of the lowest concentrations of female-authored papers—but Economic History has one of the highest (Labour Economics falls between the two). Of course, Economic History papers are still overwhelmingly—as in 74 percent—penned just by men. But given the readability gap is present in subfields with both above- and below-average rates of sole female authorship, women may need to be better writers even where more of them publish.

## E Section 3.2, suplemental output

### *E.1 Table 4, full output (first and final columns)*

Table E.1 displays coefficients from estimating Equation (1) using OLS. The first row displays coefficients on working paper score ($R_{jW}$); the second row shows the coefficient on female ratio ($\beta_{1P}$), which is also shown in the first column of Table 4. Remaining rows present estimated coefficients on the other (non-fixed effects) control variables: max. $t_5$ (author seniority), max. $T_5$ (author prominence), number of citations (asinh) and a dummy variable equal to one if article $j$ is authored by at least one native English speaker.

Similarly, Table E.2 displays coefficients from estimating Equation (2). The coefficient on female ratio corresponds to estimates presented in the final column of Table 4.

As discussed in Section 3.2.3, we do not observe the citations papers would have received had they not undergone peer review. Nevertheless, Table E.1 suggests a negative and marginally significant relationship between published readability *conditional* on draft readability; Table E.2 suggests a negative relationship or no relationship between citations and the readability improvements authors make *during* peer review.[31] Thus, they tentatively point toward women being asked to makes changes to their papers that do not ultimately improve their underlying quality.

Table E.1: Table 4 (first column), full output

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| $R_{jW}$ | 0.834*** | 0.756*** | 0.774*** | 0.791*** | 0.841*** |
| | (0.022) | (0.037) | (0.036) | (0.028) | (0.016) |
| Female ratio | 1.307** | 0.515*** | 0.513*** | 0.304** | 0.177*** |
| | (0.575) | (0.174) | (0.185) | (0.128) | (0.052) |
| $N_j$ | 0.204 | 0.084 | 0.110 | 0.073* | 0.007 |
| | (0.175) | (0.063) | (0.070) | (0.043) | (0.013) |
| Max. $t_5$ | −0.006 | 0.001 | 0.003 | 0.002 | −0.005 |
| | (0.058) | (0.014) | (0.015) | (0.009) | (0.003) |
| Max. $T_5$ | 0.011 | 0.001 | 0.000 | 0.000 | 0.003 |
| | (0.045) | (0.009) | (0.011) | (0.007) | (0.002) |
| No. citations (asinh) | −0.334* | −0.066* | −0.071 | −0.057* | −0.007 |
| | (0.180) | (0.039) | (0.046) | (0.030) | (0.015) |
| Native speaker | −0.238 | 0.000 | 0.011 | −0.021 | −0.039 |
| | (0.422) | (0.141) | (0.185) | (0.113) | (0.028) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year×Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 1,709 NBER working papers; 1,707 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles. Coefficients from OLS regression of Equation (1). First row is the coefficient on $R_{jW}$; second row is $\beta_{1P}$, and corresponds to results presented in the first column of Table 4. Coefficients on quality controls (citation counts (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority)) also shown. Standard errors clustered on editor (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[31] The coefficients on citations is negative and marginally significant ***only*** when controlling for $R_{jW}$ or using at the change in citations as the dependant variable. Otherwise, readability positively correlates with both working paper and published paper readability (results available on request; see also Appendix B.3).

TABLE E.2: Table 4 (final column), full output

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio | 0.941 | 0.437** | 0.413** | 0.237* | 0.126** |
| | (0.597) | (0.189) | (0.197) | (0.122) | (0.051) |
| $N_j$ | 0.254 | 0.096 | 0.116* | 0.080* | 0.010 |
| | (0.227) | (0.062) | (0.068) | (0.046) | (0.018) |
| Max. $t_5$ | −0.017 | 0.001 | 0.005 | 0.002 | −0.008* |
| | (0.059) | (0.017) | (0.020) | (0.013) | (0.005) |
| Max. $T_5$ | 0.019 | 0.001 | −0.001 | 0.000 | 0.005 |
| | (0.036) | (0.011) | (0.012) | (0.008) | (0.003) |
| No. citations (asinh) | −0.351 | −0.073 | −0.088 | −0.067 | −0.007 |
| | (0.224) | (0.059) | (0.069) | (0.044) | (0.017) |
| Native speaker | −0.223 | 0.061 | 0.041 | −0.021 | −0.020 |
| | (0.458) | (0.156) | (0.172) | (0.106) | (0.032) |
| Constant | 0.896 | 0.271 | 0.316 | 0.271 | −0.054 |
| | (1.147) | (0.304) | (0.327) | (0.194) | (0.087) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year×Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Sample 1,709 NBER working papers; 1,707 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles. Coefficients from OLS regression of Equation (2). First row corresponds to results presented in the final column of Table 4. Coefficients on quality controls (citation counts (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority)) also shown. Standard errors clustered on editor (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *E.2 Table 4, accounting for field*

As argued in Section 3.2 using the change in score as the dependent variable washes out any impact field may have on readability.[32] Moreover, these results—reported in the final column of Table 4—are almost identical to FGLS estimates—shown in the penultimate column—suggesting the latter are not biased by excluding them, either.

For added robustness, however, I include them here. Table E.3 replicates the analysis including dummy variables for each primary *JEL* category. As expected, figures are similar to—but standard errors somewhat higher than—those presented in Table 4.

Table E.3: Table 4, FGLS estimates controlling for *JEL* category

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.27** | 2.74*** | 3.58*** | 0.84 | 0.84 |
| | (0.57) | (1.03) | (1.17) | (0.59) | (0.60) |
| Flesch-Kincaid | 0.54*** | 0.46** | 0.88*** | 0.42** | 0.42** |
| | (0.18) | (0.23) | (0.30) | (0.20) | (0.20) |
| Gunning Fog | 0.49*** | 0.53** | 0.90*** | 0.37* | 0.37* |
| | (0.17) | (0.23) | (0.32) | (0.22) | (0.22) |
| SMOG | 0.28** | 0.38*** | 0.58*** | 0.20 | 0.20 |
| | (0.11) | (0.15) | (0.19) | (0.13) | (0.14) |
| Dale-Chall | 0.14*** | 0.32*** | 0.41*** | 0.10* | 0.10* |
| | (0.05) | (0.10) | (0.10) | (0.05) | (0.05) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |
| *JEL* (primary) effects | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,505 NBER working papers; 1,503 published articles. Estimates exclude 198 pre-internet double-blind reviewed articles (see ??). Columns display estimates identical to those in Table 4, except that fixed effects for primary *JEL* categories are included in all specifications. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[32] As long as field only impacts the readability of a paper when it is first drafted—*e.g.*, if concepts in certain areas are easier to explain—then the *change* in readability between versions is independent of it.

### *E.3 Semi-blind review*

Table 5 suggests double-blind review may have successfully reduced peer review's impact on the gender readability gap *before* the internet. Unfortunately, it has been less effective *after* the internet. I dropped NBER–published article pairs published pre-internet (*i.e.*, before Google's incorporation in 1998) and replicated Table 5 with $\text{Blind}_j$ equal to 1 if article $j$ was subjected to an official policy of double-blind review after the internet. The results, presented in Table E.4, suggests a positive gender readability gap in both samples. If anything, blinded peer review coupled with an easy alternative for determining authors' identities seems to exacerbate gender differences.

Table E.4: The impact of double-blind review after the internet

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Blind post-internet | 1.15 | 0.62** | 0.61* | 0.42* | 0.02 |
| | (1.15) | (0.31) | (0.36) | (0.24) | (0.13) |
| Non-blind | 0.74 | 0.28 | 0.22 | 0.08 | 0.15** |
| | (0.94) | (0.30) | (0.32) | (0.19) | (0.07) |
| Difference | 0.41 | 0.33 | 0.38 | 0.34 | −0.13 |
| | (1.60) | (0.45) | (0.53) | (0.34) | (0.17) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 1,380 NBER working papers; 1,378 published articles. Table replicates Table 5 with $\text{Blind}_j$ equal to 1 if article $j$ was subjected to an official policy of double-blind review after the internet. (NBER–published article pairs published pre-internet are dropped.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Editors knew submitting authors' identities—and therefore genders—both before and after the internet as well as under single- and double-blind review. Thus, the reversed gap in double-blind review pre-internet (Table 5) and positive gap post-internet (Table E.4) suggest bias from referees—as opposed to editors—may drive observed gender differences in readability.[33] Nevertheless, this conclusion is based on noisy (often insignificant) estimates. Please make it with caution.

[33] Many thanks to an anonymous referee for suggested this idea.

### *E.4 Time between working paper release and journal submission*

Figure E.1 displays a histogram of the length of time between a working paper's release and submission to *Econometrica*. It suggests most manuscripts are submitted to peer review at the same time or *before* they are released as NBER Working Papers. This is especially true of female-authored manuscripts.[34] Assuming similar submission-release patterns at *AER*, *JPE* and *QJE*, timing independence appears to be violated in only a small number of predominately male-authored papers.[35]

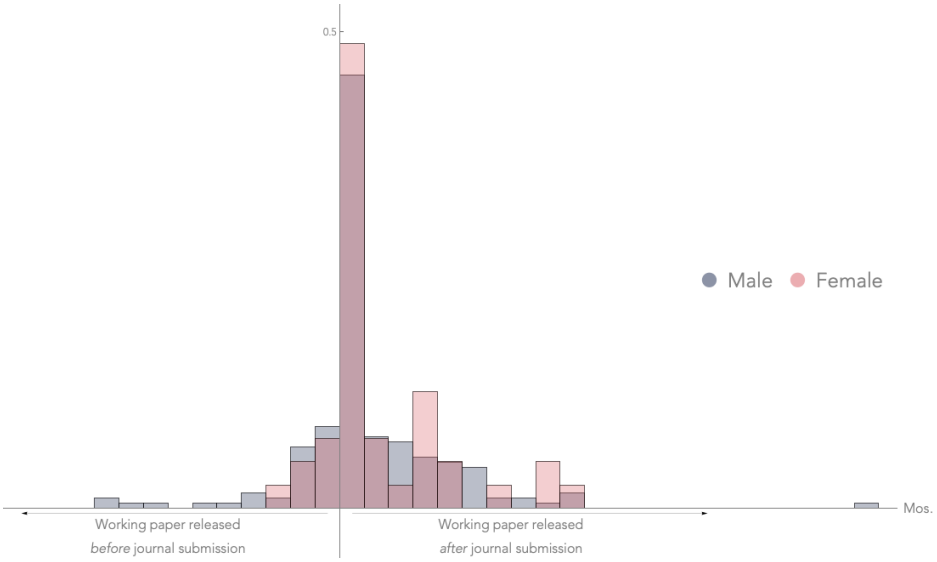


FIGURE E.1: Distribution of months between NBER release and journal submission

*Notes*. Sample 228 articles published in *Econometrica*. Pink represents papers with at least one female co-author (41 articles); blue are papers with no female co-authors (187 articles). Figure shows the distribution of the time difference (in months) between a paper's release as an NBER Working Paper and its submission to *Econometrica* (where it is eventually published). Observations on the right-hand-side of the $y$-axis were submitted to peer review first and released as working papers second; observations on the left-hand-side of the $y$-axis were released as working papers first and submitted to peer review second.

[34] Only 15 and 21 percent of female- and male-authored papers, respectively, were submitted to *Econometrica* after previously being released as an NBER Working Paper.

[35] Additionally, most drafts have *already* been widely circulated prior to NBER Working Paper release. Average acknowledgment length in NBER Working Papers is 133 words. Most authors thank at least one person for comments—indeed, the vast majority thank several—and mention having previously presented the research in conferences and seminars. Combined with evidence from Figure E.1, this suggests that gender differences in one's propensity to receive non-peer-review feedback only affects working paper readability and thus should not bias the results presented in Table 4.

### *E.5 Abstract word limits*

I attribute the change in readability between draft and final versions of a paper to the peer review process.[36] Yet NBER working paper abstracts can be of any length while abstracts published in *Econometrica* and *AER* cannot—they are restricted to 150 and 100 words, respectively. Observed readability gaps could consequently result from gender differences in how authors conform to these limits.

To test this hypothesis, I replicated the analysis described Section 3.2.3 (and shown in Table 4) on the subset of articles with draft abstracts below the official minimum word limit of the journals in which they were eventually published. Results are shown in Table E.5. Despite dropping about 40 percent of observations, coefficient magnitudes are similar to those reported in Table 4; standard errors are somewhat larger.[37]

Table E.5: Table 4, draft abstracts below official word limits

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 0.82 | 2.33 | 2.78* | 0.45 | 0.45 |
| | (0.86) | (1.47) | (1.56) | (0.83) | (0.86) |
| Flesch-Kincaid | 0.53** | 0.07 | 0.59* | 0.52* | 0.52* |
| | (0.26) | (0.34) | (0.33) | (0.27) | (0.28) |
| Gunning Fog | 0.55** | 0.23 | 0.73** | 0.50* | 0.50* |
| | (0.24) | (0.38) | (0.34) | (0.26) | (0.27) |
| SMOG | 0.27* | 0.23 | 0.45** | 0.22 | 0.22 |
| | (0.15) | (0.26) | (0.22) | (0.16) | (0.16) |
| Dale-Chall | 0.23*** | 0.33*** | 0.50*** | 0.17** | 0.17** |
| | (0.08) | (0.12) | (0.11) | (0.07) | (0.08) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,067 NBER working papers; 1,065 published articles. Estimates are identical to those in Table 4, except that the sample includes only papers with an NBER abstract below the official minimum word limit of the journal in which it was eventually published. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[36] See Section 3.2.4 for a more detailed discussion and justification of the assumptions underpinning this claim.

[37] Results are similar if I also include a control for the number of words in the working paper version of the abstract (available on request).

## F Section 3.3, supplemental output

### *F.1 Authors' average readability scores for their first, mean and final papers*

Table F.1 displays authors' average readability scores for their first, mean and final top-four papers. Grade-level scores (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (see Section 2.2). Sample excludes authors with fewer than three publications.

As their careers advance, women do write more clearly: their average readability scores are 1–5 percent higher than the readability of their first papers; their latest papers 1–7 percent. For a man, however, his average and last paper may be more poorly written than the first.

Table F.1: Average first, mean and final top-four paper scores

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| **Average first paper score** | | | | | |
| Men | 39.37 | −13.77 | −17.54 | −15.35 | −11.00 |
| | (0.31) | (0.07) | (0.08) | (0.06) | (0.03) |
| Women | 39.20 | −13.81 | −17.36 | −15.18 | −11.00 |
| | (1.15) | (0.24) | (0.29) | (0.21) | (0.10) |
| **Average mean score** | | | | | |
| Men | 39.59 | −13.69 | −17.41 | −15.26 | −11.02 |
| | (0.19) | (0.04) | (0.05) | (0.03) | (0.02) |
| Women | 41.20 | −13.36 | −16.92 | −14.92 | −10.91 |
| | (0.72) | (0.15) | (0.19) | (0.14) | (0.07) |
| **Average final paper score** | | | | | |
| Men | 39.54 | −13.71 | −17.41 | −15.24 | −11.08 |
| | (0.33) | (0.08) | (0.09) | (0.06) | (0.03) |
| Women | 41.99 | −13.10 | −16.58 | −14.66 | −10.90 |
| | (1.06) | (0.21) | (0.25) | (0.18) | (0.11) |

*Notes*. Sample 1,675 authors; includes only authors with three or more publications. Figures are average readability scores for authors' first, mean and last published articles. Grade-level scores have been multiplied by negative one (see ??). Standard errors in parentheses.

### *F.2 Table 6, tests of coefficient equality*

Table F.2 tests equality of coefficients in each column of Table 6. It rejects equality between coefficients in the first and third columns at $p < 0.01$ for the Flesch Reading Ease, Flesch-Kincaid and SMOG scores and $p < 0.05$ for the Gunning Fog and Dale-Chall scores.

TABLE F.2: Table F.2, equality test statistics

| | $t_4 = 1$ vs. 2 | 1 vs. 3 | 1 vs. 4–5 | 1 vs. $\geq 6$ | 2 vs. 3 |
|---|---|---|---|---|---|
| Flesch Reading Ease | 1.172 | 11.204 | 1.670 | 0.681 | 6.495 |
| Flesch-Kincaid | 0.074 | 11.996 | 1.672 | 0.744 | 9.607 |
| Gunning Fog | 0.181 | 11.394 | 2.036 | 0.804 | 8.764 |
| SMOG | 0.338 | 10.536 | 2.052 | 1.006 | 7.014 |
| Dale-Chall | 0.185 | 4.787 | 1.698 | 1.846 | 3.293 |

*Notes.* $\chi^2$ test statistics from Wald tests of $\beta_1$ (Equation (H.1)) equality across estimation results in Table 6.

### *F.3 Co-variate balance*

Table F.3 compares co-variate balance pre- and post-match.[38] The first column displays averages for the 121 female authors with at least three publications in the data. The first column of the first panel ("Pre-match means") displays corresponding averages for the 1,554 male authors with three or more publications. The first column of the second panel ("Post-match means") displays (weighted) averages for the 110 male authors matched with a female author. Table F.4, Table F.5 and Table F.6 compare co-variate balance when restricted to matched pairs with $\underline{D}_{ik} \neq 0$.

Gender differences are smaller post-match; $t$-statistics are likewise closer to zero. Moreover, co-variates remain well balanced—and resemble averages in the matched sample—in both $\underline{D}_{ik} > 0$ (discrimination against women) and $\underline{D}_{ik} < 0$ (discrimination against men) samples (not shown).

[38] Matches were generated in Stata using `psmatch2` (Leuven and Sianesi, 2003).

Table F.3: Pre- and post-matching summary statistics

| | | Pre-match means | | | Post-match means | | |
|---|---|---|---|---|---|---|---|
| | Women | Men | Difference | $t$ | Men | Difference | $t$ |
| $t = 1$ inst. rank | 18.47 | 15.26 | 3.22 | 1.85 | 15.26 | 1.89 | 0.79 |
| Max. citations | 406.33 | 267.07 | 139.27 | 1.78 | 267.07 | −73.52 | −2.00 |
| **Fraction of articles per decade** | | | | | | | |
| 1950–59 | 0.01 | 0.00 | 0.01 | 1.57 | 0.00 | 0.00 | |
| 1960–69 | 0.04 | 0.00 | 0.04 | 2.87 | 0.00 | 0.00 | −0.12 |
| 1970–79 | 0.11 | 0.01 | 0.09 | 4.72 | 0.01 | 0.00 | 0.10 |
| 1980–89 | 0.18 | 0.08 | 0.10 | 4.37 | 0.08 | 0.01 | 0.55 |
| 1990–99 | 0.21 | 0.19 | 0.02 | 1.00 | 0.19 | 0.00 | −0.07 |
| 2000–09 | 0.26 | 0.41 | −0.15 | −5.90 | 0.41 | −0.01 | −0.22 |
| 2010–15 | 0.20 | 0.31 | −0.11 | −4.19 | 0.31 | 0.00 | −0.09 |
| **Fraction of articles per journal** | | | | | | | |
| *AER* | 0.25 | 0.39 | −0.14 | −5.54 | 0.39 | −0.01 | −0.29 |
| *Econometrica* | 0.34 | 0.17 | 0.17 | 5.12 | 0.17 | 0.07 | 1.77 |
| *JPE* | 0.24 | 0.18 | 0.07 | 2.62 | 0.18 | 0.00 | −0.10 |
| *QJE* | 0.17 | 0.27 | −0.10 | −4.79 | 0.27 | −0.06 | −1.85 |
| **Number of articles per *JEL* code** | | | | | | | |
| A General | 0.02 | 0.04 | −0.02 | −1.59 | 0.04 | 0.00 | 0.00 |
| B Methodology | 0.02 | 0.00 | 0.02 | 1.44 | 0.00 | 0.00 | |
| C Quant. methods | 0.81 | 0.64 | 0.17 | 1.03 | 0.64 | −0.07 | −0.44 |
| D Microeconomics | 1.79 | 1.64 | 0.15 | 0.68 | 1.64 | −0.03 | −0.15 |
| E Macroeconomics | 0.62 | 0.58 | 0.04 | 0.35 | 0.58 | −0.07 | −0.55 |
| F International | 0.31 | 0.39 | −0.08 | −0.85 | 0.39 | −0.09 | −0.75 |
| G Finance | 0.52 | 0.60 | −0.07 | −0.67 | 0.60 | −0.13 | −0.90 |
| H Public | 0.36 | 0.45 | −0.10 | −1.09 | 0.45 | −0.17 | −2.08 |
| I Health, welfare, edu | 0.34 | 0.88 | −0.53 | −5.40 | 0.88 | −0.36 | −2.01 |
| J Labour | 0.76 | 1.26 | −0.49 | −3.39 | 1.26 | −0.41 | −2.18 |
| K Law and econ | 0.14 | 0.20 | −0.06 | −1.14 | 0.20 | −0.07 | −1.03 |
| L Industrial org | 0.57 | 0.73 | −0.16 | −1.47 | 0.73 | −0.23 | −1.78 |
| M Marketing/acct | 0.13 | 0.17 | −0.04 | −0.93 | 0.17 | −0.04 | −0.65 |
| N Economic history | 0.14 | 0.29 | −0.15 | −2.74 | 0.29 | −0.09 | −0.97 |
| O Development | 0.52 | 0.86 | −0.34 | −2.60 | 0.86 | −0.36 | −2.21 |
| P Economic systems | 0.09 | 0.08 | 0.01 | 0.22 | 0.08 | −0.02 | −0.57 |
| Q Agri., environment | 0.12 | 0.18 | −0.06 | −1.20 | 0.18 | −0.09 | −1.51 |
| R Regional, transport | 0.16 | 0.17 | −0.01 | −0.16 | 0.17 | −0.12 | −2.67 |
| Z Special topics | 0.10 | 0.16 | −0.06 | −1.50 | 0.16 | −0.01 | −0.14 |

*Notes.* Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics (1,554 female authors, 121 male authors). Second panel shows post-match summary statistics (109 male authors). $t$-values for differences reported in each panel's final column.

TABLE F.4: Co-variate post-match balance when $\underline{D}_{ik} \neq 0$

| | Flesch Reading Ease | | | | Flesch Kincaid | | | |
|---|---|---|---|---|---|---|---|---|
| | Discrimination | | | | Discrimination | | | |
| | Against women | Against men | Difference | $t$ | Against women | Against men | Difference | $t$ |
| $t = 1$ inst. rank | 15.55 | 17.01 | −1.46 | −0.52 | 16.65 | 15.97 | 0.68 | 0.24 |
| Max. citations | 204.72 | 232.66 | −27.94 | −0.89 | 210.06 | 266.92 | −56.86 | −1.31 |
| **Fraction of articles per decade** | | | | | | | | |
| 1950–59 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | |
| 1960–69 | 0.00 | 0.00 | 0.00 | −0.18 | 0.00 | 0.00 | 0.00 | −0.20 |
| 1970–79 | 0.02 | 0.02 | 0.00 | −0.01 | 0.01 | 0.01 | 0.00 | −0.08 |
| 1980–89 | 0.07 | 0.08 | −0.01 | −0.26 | 0.09 | 0.09 | 0.00 | −0.01 |
| 1990–99 | 0.18 | 0.20 | −0.02 | −0.50 | 0.17 | 0.18 | −0.01 | −0.31 |
| 2000–09 | 0.43 | 0.43 | −0.01 | −0.12 | 0.41 | 0.40 | 0.00 | 0.11 |
| 2010–15 | 0.30 | 0.27 | 0.03 | 0.72 | 0.32 | 0.31 | 0.01 | 0.20 |
| **Fraction of articles per journal** | | | | | | | | |
| *AER* | 0.38 | 0.39 | −0.01 | −0.34 | 0.40 | 0.40 | 0.00 | 0.01 |
| *Econometrica* | 0.23 | 0.16 | 0.07 | 1.47 | 0.24 | 0.19 | 0.05 | 0.98 |
| *JPE* | 0.16 | 0.19 | −0.03 | −0.95 | 0.14 | 0.17 | −0.03 | −1.02 |
| *QJE* | 0.23 | 0.26 | −0.02 | −0.61 | 0.22 | 0.24 | −0.02 | −0.43 |
| **Fraction of articles per *JEL* code** | | | | | | | | |
| A General | 0.04 | 0.05 | 0.00 | −0.12 | 0.02 | 0.02 | 0.00 | −0.11 |
| B Methodology | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | |
| C Quant. methods | 0.49 | 0.41 | 0.08 | 0.46 | 0.53 | 0.55 | −0.02 | −0.10 |
| D Microeconomics | 1.54 | 1.55 | −0.02 | −0.07 | 1.60 | 1.59 | 0.01 | 0.05 |
| E Macroeconomics | 0.43 | 0.49 | −0.06 | −0.44 | 0.42 | 0.48 | −0.06 | −0.42 |
| F International | 0.26 | 0.38 | −0.12 | −0.81 | 0.26 | 0.33 | −0.07 | −0.48 |
| G Finance | 0.45 | 0.49 | −0.04 | −0.23 | 0.47 | 0.60 | −0.14 | −0.75 |
| H Public | 0.31 | 0.39 | −0.08 | −0.83 | 0.36 | 0.40 | −0.03 | −0.30 |
| I Health, welfare, edu | 0.70 | 0.88 | −0.18 | −0.80 | 0.77 | 0.81 | −0.04 | −0.19 |
| J Labour | 0.84 | 1.08 | −0.24 | −1.14 | 0.87 | 1.09 | −0.22 | −1.03 |
| K Law and econ | 0.12 | 0.16 | −0.05 | −0.67 | 0.12 | 0.21 | −0.08 | −1.07 |
| L Industrial org | 0.58 | 0.66 | −0.08 | −0.49 | 0.55 | 0.72 | −0.17 | −1.08 |
| M Marketing/acct | 0.12 | 0.13 | −0.01 | −0.16 | 0.11 | 0.13 | −0.01 | −0.20 |
| N Economic history | 0.22 | 0.36 | −0.15 | −1.18 | 0.19 | 0.29 | −0.10 | −0.89 |
| O Development | 0.40 | 0.62 | −0.23 | −1.64 | 0.40 | 0.62 | −0.22 | −1.59 |
| P Economic systems | 0.05 | 0.09 | −0.04 | −0.76 | 0.06 | 0.10 | −0.04 | −0.75 |
| Q Agri., environment | 0.06 | 0.14 | −0.08 | −1.37 | 0.04 | 0.09 | −0.05 | −1.23 |
| R Regional, transport | 0.06 | 0.12 | −0.05 | −1.10 | 0.07 | 0.16 | −0.09 | −1.73 |
| Z Special topics | 0.17 | 0.18 | 0.00 | −0.06 | 0.15 | 0.17 | −0.03 | −0.42 |

*Notes.* Sample restricted to authors with three or more publications. Panels show post-match summary statistics for pairs in which $\underline{D}_{ik} \neq 0$. $t$-values for differences reported in each panel's final column.

TABLE F.5: Co-variate post-match balance when $\underline{D}_{ik} \neq 0$

| | Gunning Fog | | | | SMOG | | | |
|---|---|---|---|---|---|---|---|---|
| | Discrimination | | | | Discrimination | | | |
| | Against women | Against men | Difference | $t$ | Against women | Against men | Difference | $t$ |
| $t = 1$ inst. rank | 16.37 | 18.10 | −1.73 | −0.59 | 14.84 | 16.93 | −2.09 | −0.71 |
| Max. citations | 193.79 | 245.42 | −51.62 | −1.55 | 198.31 | 240.97 | −42.66 | −1.23 |
| **Fraction of articles per decade** | | | | | | | | |
| 1950–59 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | |
| 1960–69 | 0.00 | 0.00 | 0.00 | −0.20 | 0.00 | 0.00 | 0.00 | −0.13 |
| 1970–79 | 0.02 | 0.02 | 0.00 | −0.18 | 0.02 | 0.02 | 0.00 | 0.08 |
| 1980–89 | 0.10 | 0.09 | 0.00 | 0.07 | 0.09 | 0.08 | 0.01 | 0.39 |
| 1990–99 | 0.16 | 0.17 | −0.01 | −0.26 | 0.16 | 0.18 | −0.02 | −0.55 |
| 2000–09 | 0.40 | 0.39 | 0.00 | 0.05 | 0.40 | 0.42 | −0.02 | −0.38 |
| 2010–15 | 0.32 | 0.31 | 0.01 | 0.17 | 0.33 | 0.31 | 0.02 | 0.48 |
| **Fraction of articles per journal** | | | | | | | | |
| *AER* | 0.40 | 0.38 | 0.01 | 0.25 | 0.39 | 0.40 | −0.01 | −0.19 |
| *Econometrica* | 0.22 | 0.21 | 0.02 | 0.31 | 0.21 | 0.17 | 0.04 | 0.88 |
| *JPE* | 0.15 | 0.18 | −0.03 | −0.79 | 0.16 | 0.20 | −0.04 | −1.19 |
| *QJE* | 0.23 | 0.23 | 0.00 | −0.01 | 0.24 | 0.23 | 0.01 | 0.25 |
| **Fraction of articles per *JEL* code** | | | | | | | | |
| A General | 0.02 | 0.02 | 0.00 | −0.12 | 0.03 | 0.03 | 0.00 | −0.01 |
| B Methodology | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | |
| C Quant. methods | 0.46 | 0.51 | −0.05 | −0.31 | 0.44 | 0.45 | −0.01 | −0.04 |
| D Microeconomics | 1.43 | 1.45 | −0.02 | −0.09 | 1.43 | 1.52 | −0.09 | −0.37 |
| E Macroeconomics | 0.49 | 0.45 | 0.04 | 0.28 | 0.52 | 0.45 | 0.07 | 0.48 |
| F International | 0.29 | 0.26 | 0.03 | 0.28 | 0.27 | 0.23 | 0.04 | 0.36 |
| G Finance | 0.42 | 0.52 | −0.10 | −0.58 | 0.48 | 0.48 | −0.01 | −0.03 |
| H Public | 0.40 | 0.42 | −0.02 | −0.16 | 0.41 | 0.40 | 0.01 | 0.06 |
| I Health, welfare, edu | 0.70 | 0.85 | −0.15 | −0.66 | 0.61 | 1.02 | −0.41 | −1.76 |
| J Labour | 0.93 | 1.14 | −0.21 | −0.99 | 0.99 | 1.10 | −0.11 | −0.51 |
| K Law and econ | 0.09 | 0.16 | −0.07 | −1.15 | 0.11 | 0.16 | −0.05 | −0.73 |
| L Industrial org | 0.42 | 0.56 | −0.14 | −0.97 | 0.56 | 0.51 | 0.05 | 0.35 |
| M Marketing/acct | 0.09 | 0.09 | 0.00 | 0.00 | 0.10 | 0.09 | 0.01 | 0.20 |
| N Economic history | 0.28 | 0.41 | −0.13 | −1.01 | 0.30 | 0.41 | −0.12 | −0.88 |
| O Development | 0.40 | 0.59 | −0.19 | −1.48 | 0.50 | 0.49 | 0.01 | 0.04 |
| P Economic systems | 0.06 | 0.09 | −0.03 | −0.56 | 0.05 | 0.08 | −0.04 | −0.71 |
| Q Agri., environment | 0.06 | 0.10 | −0.04 | −0.83 | 0.07 | 0.10 | −0.04 | −0.65 |
| R Regional, transport | 0.05 | 0.14 | −0.09 | −1.96 | 0.09 | 0.13 | −0.04 | −0.71 |
| Z Special topics | 0.13 | 0.14 | −0.01 | −0.08 | 0.15 | 0.15 | 0.00 | −0.02 |

*Notes.* Sample restricted to authors with three or more publications. Panels show post-match summary statistics for pairs in which $\underline{D}_{ik} \neq 0$. $t$-values for differences reported in each panel's final column.

TABLE F.6: Co-variate post-match balance when $\underline{D}_{ik} \neq 0$

| | Dale-Chall | | | |
|---|---|---|---|---|
| | Discrimination | | | |
| | Against women | Against men | Difference | $t$ |
| $t = 1$ inst. rank | 14.54 | 16.70 | −2.17 | −0.83 |
| Max. citations | 214.67 | 278.53 | −63.85 | −1.41 |
| **Fraction of articles per decade** | | | | |
| 1950–59 | 0.00 | 0.00 | 0.00 | |
| 1960–69 | 0.00 | 0.00 | 0.00 | −0.15 |
| 1970–79 | 0.02 | 0.02 | 0.00 | 0.04 |
| 1980–89 | 0.08 | 0.09 | 0.00 | −0.08 |
| 1990–99 | 0.16 | 0.19 | −0.03 | −0.79 |
| 2000–09 | 0.42 | 0.43 | −0.01 | −0.25 |
| 2010–15 | 0.31 | 0.27 | 0.04 | 0.95 |
| **Fraction of articles per journal** | | | | |
| *AER* | 0.38 | 0.38 | 0.00 | −0.06 |
| *Econometrica* | 0.22 | 0.16 | 0.07 | 1.47 |
| *JPE* | 0.18 | 0.20 | −0.02 | −0.61 |
| *QJE* | 0.22 | 0.26 | −0.04 | −1.13 |
| **Fraction of articles per *JEL* code** | | | | |
| A General | 0.04 | 0.04 | 0.00 | −0.06 |
| B Methodology | 0.00 | 0.00 | 0.00 | |
| C Quant. methods | 0.45 | 0.36 | 0.09 | 0.72 |
| D Microeconomics | 1.74 | 1.57 | 0.17 | 0.68 |
| E Macroeconomics | 0.64 | 0.55 | 0.09 | 0.54 |
| F International | 0.29 | 0.30 | 0.00 | −0.01 |
| G Finance | 0.51 | 0.54 | −0.03 | −0.18 |
| H Public | 0.42 | 0.35 | 0.07 | 0.67 |
| I Health, welfare, edu | 0.51 | 0.97 | −0.46 | −2.15 |
| J Labour | 0.95 | 1.22 | −0.27 | −1.15 |
| K Law and econ | 0.19 | 0.22 | −0.03 | −0.36 |
| L Industrial org | 0.64 | 0.57 | 0.07 | 0.47 |
| M Marketing/acct | 0.13 | 0.14 | −0.02 | −0.24 |
| N Economic history | 0.27 | 0.35 | −0.08 | −0.63 |
| O Development | 0.58 | 0.75 | −0.17 | −0.91 |
| P Economic systems | 0.07 | 0.11 | −0.04 | −0.63 |
| Q Agri., environment | 0.08 | 0.14 | −0.06 | −0.88 |
| R Regional, transport | 0.13 | 0.12 | 0.01 | 0.10 |
| Z Special topics | 0.16 | 0.19 | −0.03 | −0.42 |

*Notes.* Sample restricted to authors with three or more publications. Panels show post-match summary statistics for pairs in which $\underline{D}_{ik} \neq 0$. $t$-values for differences reported in each panel's final column.

### *F.4 List of authors in each matched pair*

Table F.7: Matched pairs

| Matched pairs | | Matched pairs | |
|---|---|---|---|
| Female | Male | Female | Male |
| Abraham, Katharine G. | Kahn, Charles M. | Kuziemko, Ilyana | Deming, David J. |
| Admati, Anat R. | Huang, Chi-Fu | La Ferrara, Eliana | Krebs, Tom |
| Amiti, Mary | Broda, Christian | Landes, Elisabeth M. | Carlton, Dennis W. |
| Anderson, Siwan | Baland, Jean-Marie | Levy, Gilat | Razin, Ronny |
| Ashraf, Nava | Avery, Christopher N. | Lewis, Karen K. | Backus, David K. |
| Athey, Susan | Haile, Philip A. | Li, Wei | Roland, Gérard |
| Baicker, Katherine | Shafir, Eldar | Lleras-Muney, Adriana | Kessler, Daniel P. |
| Bailey, Martha J. | Paserman, M. Daniele | Løken, Katrine Vellesen | Mogstad, Magne |
| Bandiera, Oriana | Rasul, Imran | Madrian, Brigitte C. | Lee, David S. |
| Barwick, Panle Jia | Winston, Clifford | Maestas, Nicole | Bettinger, Eric P. |
| Baxter, Marianne | Backus, David K. | Malmendier, Ulrike | Agarwal, Sumit |
| Bedard, Kelly | Lefgren, Lars | Matzkin, Rosa L. | Hahn, Jinyong |
| Bertrand, Marianne | Mullainathan, Sendhil | McConnell, Sheena | LaLonde, Robert J. |
| Black, Sandra E. | Kessler, Daniel P. | McGrattan, Ellen R. | Williams, Noah |
| Blank, Rebecca M. | Laband, David N. | Meyer, Margaret A. | Holtz-Eakin, Douglas |
| Boustan, Leah Platt | Abramitzky, Ran | Molinari, Francesca | Hansen, Peter Reinhard |
| Brown, Jennifer | Vogel, Jonathan | Moser, Petra | Sunde, Uwe |
| Busse, Meghan R. | Zettelmeyer, Florian | Nakamura, Emi | Steinsson, Jón |
| Case, Anne C. | Scholz, John Karl | Ng, Serena | Muller, Ulrich K. |
| Casella, Alessandra | Snyder, James M. (Jr.) | Niederle, Muriel | Wolfers, Justin |
| Chen, Xiaohong | Hahn, Jinyong | Oster, Emily | Fang, Hanming |
| Chen, Yan | Lange, Andreas | Pande, Rohini | Dean, Mark |
| Chevalier, Judith A. | Lamont, Owen A. | Paxson, Christina H. | Boldrin, Michele |
| Chichilnisky, Graciela | Engers, Maxim | Perrigne, Isabelle | Schmedders, Karl |
| Correia, Isabel | Leeper, Eric M. | Piazzesi, Monika | Schneider, Martin |
| Costa, Dora L. | Kahn, Matthew E. | Qian, Nancy | Ok, Efe A. |
| Cropper, Maureen L. | Halvorsen, Robert | Quinzii, Martine | Magill, Michael J. P. |
| Currie, Janet | Lavy, Victor | Ramey, Valerie A. | Salanié, Bernard |
| Dafny, Leemore S. | Kolstad, Jonathan T. | Reinganum, Jennifer F. | Daughety, Andrew F. |
| De Nardi, Mariacristina | Silverman, Dan | Reinhart, Carmen M. | Taylor, Alan M. |
| Demange, Gabrielle | Anderson, Robert M. | Rey, Hélène | Jeanne, Olivier |
| Duflo, Esther | Burgess, Robin | Romer, Christina D. | Williams, John C. |
| Dupas, Pascaline | Urquiola, Miguel | Rose-Ackerman, Susan | Miyazaki, Hajime |
| Dynan, Karen E. | Ljungqvist, Lars | Rose, Nancy L. | Snyder, James M. (Jr.) |
| Eberly, Janice C. | Sunder, Shyam | Rosenblat, Tanya S. | Möbius, Markus M. |
| Eckel, Catherine C. | Dufwenberg, Martin | Rouse, Cecilia Elena | Fishman, Arthur |
| Edlund, Lena | Smith, Jeffrey | Sapienza, Paola | Wacziarg, Romain |
| Eyigungor, Burcu | Kaboski, Joseph P. | Schennach, Susanne M. | Guggenberger, Patrik |
| Fan, Yanqin | Rahbek, Anders | Schmitt-Grohé, Stephanie | Leeper, Eric M. |
| Fernández, Raquel | Spolaore, Enrico | Schwartz, Nancy L. | Kuga, Kiyoshi |
| Field, Erica | Donald, Stephen G. | Shannon, Chris | Safra, Zvi |
| Finkelstein, Amy | Einav, Liran | Shaw, Kathryn L. | Anderson, Simon P. |
| Flavin, Marjorie A. | Garber, Peter M. | Spier, Kathryn E. | Ausubel, Lawrence M. |
| Forges, Françoise | Hong, Chew Soo | Stokey, Nancy L. | Smith, Bruce D. |
| Fortin, Nicole M. | Hyslop, Dean R. | Tenreyro, Silvana | Lloyd-Ellis, Huw |
| Freund, Caroline | Rose, Andrew K. | Tertilt, Michèle | Doepke, Matthias |
| Fuchs-Schündeln, Nicola | Woodruff, Christopher | Tesar, Linda L. | Blonigen, Bruce A. |
| Garfinkel, Michelle R. | Bertola, Giuseppe | Thomas, Julia K. | Khan, Aubhik |
| Goldberg, Pinelopi Koujianou | Levinsohn, James A. | Todd, Petra E. | Flinn, Christopher J. |
| Goldin, Claudia D. | Abramitzky, Ran | Vissing-Jørgensen, Annette | Veronesi, Pietro |
| Gopinath, Gita | Itskhoki, Oleg | Voena, Alessandra | Sunde, Uwe |
| Griffith, Rachel | Broda, Christian | Washington, Ebonya L. | Paserman, M. Daniele |
| Guerrieri, Veronica | Khan, Aubhik | White, Lucy | Yılmaz, Bilge |
| Hanna, Rema | Foster, Andrew D. | Whited, Toni M. | Sun, Ning |
| Hastings, Justine S. | Pope, Devin G. | Williams, Heidi L. | Budish, Eric |
| Ho, Katherine | Kremer, Ilan | Wooders, Myrna Holtz | Gallant, A. Ronald |

Table F.7 (continued)

| Matched pairs | | Matched pairs | |
|---|---|---|---|
| Female | Male | Female | Male |
| Hoxby, Caroline Minter | Kessler, Daniel P. | Yariv, Leeat | Lange, Andreas |
| İmrohoroğlu, Ayşe | Casari, Marco | Yellen, Janet L. | Freeman, Richard B. |
| Jayachandran, Seema | Pop-Eleches, Cristian | Zeiler, Kathryn | van Soest, Arthur |
| Kowalski, Amanda E. | Schrimpf, Paul | Zhuravskaya, Ekaterina | Kuhn, Peter |
| Kranton, Rachel E. | Kosfeld, Michael | | |

*Notes.* Table lists the names of the matched pairs from Section 3.3.2. In each panel, female members are listed first; male members second. See Section 3.3.2 for details on the matching process.

### *F.5 $\widehat{R}_{it}$ regression output*

Table F.8 displays output from time- and gender-specific regressions used to generate $\widehat{R}_{it}$ (Equation (13)).

TABLE F.8: Regression output generating $\widehat{R}_{it}$ (Equation (13))

| | Women | | Men | |
|---|---|---|---|---|
| | $t_4 = 1$ | $t_4 = 3$ | $t_4 = 1$ | $t_4 = 3$ |
| **Flesch Reading Ease** | | | | |
| Female ratio | 1.36 | 2.99 | −6.02 | 6.34 |
| | (4.16) | (3.88) | (8.54) | (5.80) |
| Constant | 38.24*** | 41.17*** | 37.57*** | 38.02*** |
| | (3.15) | (2.47) | (1.15) | (1.22) |
| **Flesch Kincaid** | | | | |
| Female ratio | −0.13 | 0.48 | 0.34 | 2.41* |
| | (0.86) | (0.78) | (1.90) | (1.26) |
| Constant | −13.72*** | −13.33*** | −14.14*** | −14.36*** |
| | (0.65) | (0.50) | (0.25) | (0.26) |
| **Gunning Fog** | | | | |
| Female ratio | −0.30 | 1.01 | −0.88 | 2.51 |
| | (1.04) | (0.97) | (2.11) | (1.52) |
| Constant | −17.15*** | −17.22*** | −17.97*** | −18.00*** |
| | (0.79) | (0.62) | (0.28) | (0.32) |
| **SMOG** | | | | |
| Female ratio | −0.15 | 0.74 | −0.45 | 1.56 |
| | (0.76) | (0.72) | (1.46) | (1.10) |
| Constant | −15.07*** | −15.19*** | −15.72*** | −15.63*** |
| | (0.57) | (0.46) | (0.20) | (0.23) |
| **Dale-Chall** | | | | |
| Female ratio | −0.06 | 0.48 | −2.02** | 0.37 |
| | (0.35) | (0.39) | (0.82) | (0.42) |
| Constant | −10.96*** | −11.11*** | −11.11*** | −11.16*** |
| | (0.26) | (0.25) | (0.11) | (0.09) |

*Notes*. Sample 121 female authors; 109 male authors. Sample restricted to matched authors. See Section 3.3.2 for details on how matches were made. Regressions weighted by the frequency observations are used in a match. Standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *F.6 Table 7, Equation (11) and Condition 3*

Table F.9 estimates $D_{ik}$ with Equation (11). Table F.10 estimates $D_{ik}$ with a rough attempt to control for acceptance rates—it requires $T_{5i} \leq T_{5k}$ or $T_{5k} \leq T_{5i}$ before categorising matched pairs as discrimination against $i$ or $k$, respectively. Conclusions from both tables are are similar to those presented in Section 3.3.2.

TABLE F.9: $D_{ik}$, Equation (11)

| | Discrimination against women ($\underline{D}_{ik} > 0$) | | | Discrimination against men ($\underline{D}_{ik} < 0$) | | | Mean, all observations | |
|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | $N$ | Mean | S.D. | $N$ | (1) | (2) |
| Flesch Reading Ease | 9.01 | 7.44 | 60 | −5.80 | 5.85 | 20 | 3.38*** | 2.34*** |
| | | | | | | | (0.77) | (0.89) |
| Flesch Kincaid | 1.76 | 1.29 | 65 | −1.52 | 1.49 | 19 | 0.75*** | 0.61*** |
| | | | | | | | (0.16) | (0.17) |
| Gunning Fog | 2.32 | 1.82 | 62 | −1.60 | 1.80 | 21 | 0.95*** | 0.76*** |
| | | | | | | | (0.20) | (0.23) |
| SMOG | 1.82 | 1.38 | 54 | −0.94 | 1.16 | 24 | 0.61*** | 0.46*** |
| | | | | | | | (0.15) | (0.17) |
| Dale-Chall | 0.88 | 0.65 | 62 | −0.68 | 0.48 | 22 | 0.32*** | 0.23*** |
| | | | | | | | (0.08) | (0.09) |

*Notes.* Table displays estimates identical to those in Table 7, except that $\underline{D}_{ik}$ is determined by Equation (11). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE F.10: $D_{ik}$, proxying for acceptance rates (Condition 3)

| | Discrimination against women ($\underline{D}_{ik} > 0$) | | | Discrimination against men ($\underline{D}_{ik} < 0$) | | | Mean, all observations | |
|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | $N$ | Mean | S.D. | $N$ | (1) | (2) |
| Flesch Reading Ease | 13.64 | 10.99 | 35 | −7.90 | 7.22 | 14 | 3.63*** | 2.21* |
| | | | | | | | (1.18) | (1.32) |
| Flesch Kincaid | 2.77 | 2.30 | 39 | −2.33 | 2.19 | 16 | 0.74*** | 0.55* |
| | | | | | | | (0.27) | (0.28) |
| Gunning Fog | 3.25 | 2.99 | 40 | −2.44 | 2.86 | 16 | 0.91*** | 0.66* |
| | | | | | | | (0.32) | (0.35) |
| SMOG | 2.73 | 2.14 | 33 | −1.45 | 2.07 | 16 | 0.63*** | 0.42* |
| | | | | | | | (0.23) | (0.25) |
| Dale-Chall | 1.35 | 0.98 | 35 | −1.01 | 0.77 | 16 | 0.35*** | 0.21 |
| | | | | | | | (0.13) | (0.14) |

*Notes.* Table displays estimates identical to those in Table 7, except that a matched pair is categorised as discrimination against $i$ ($k$) only if $T_{5i} \leq T_{5k}$ ($T_{5k} \leq T_{5i}$) holds as well. Otherwise, Theorem 1 is inconclusive. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *F.7* $\widehat{R}_{it}$, *controlling for* JEL *category*

Table F.11 and Figure F.1 replicate the analysis in Section 3.3.2 but Equation (13) controls for primary *JEL* category. $\widehat{R}_{it}$ was reconstructed at female ratio equal to 1 for women, 0 for men and for a paper classified in *JEL* categories D (microeconomics) and J (labour and demographic economics).

TABLE F.11: $\underline{D}_{ik}$, controlling for *JEL* category

| | Discrimination against women ($\underline{D}_{ik} > 0$) | | | Discrimination against men ($\underline{D}_{ik} < 0$) | | | Mean, all observations | |
|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | $N$ | Mean | S.D. | $N$ | (1) | (2) |
| Flesch Reading Ease | 18.27 | 11.92 | 50 | −7.59 | 7.96 | 6 | 9.75*** | 9.13*** |
| | | | | | | | (2.57) | (2.72) |
| Flesch Kincaid | 3.87 | 2.71 | 45 | −1.46 | 1.05 | 8 | 1.84*** | 1.65*** |
| | | | | | | | (0.54) | (0.59) |
| Gunning Fog | 4.49 | 2.98 | 40 | −2.14 | 2.45 | 13 | 1.76*** | 1.51** |
| | | | | | | | (0.63) | (0.68) |
| SMOG | 3.22 | 2.19 | 41 | −1.81 | 1.87 | 11 | 1.30*** | 1.15** |
| | | | | | | | (0.46) | (0.49) |
| Dale-Chall | 1.80 | 1.11 | 32 | −1.09 | 0.76 | 5 | 0.61*** | 0.53** |
| | | | | | | | (0.22) | (0.23) |

*Notes.* Sample 88 matched pairs (79 and 88 distinct men and women, respectively). Table displays estimates identical to those in Table 7, except that Equation (13) includes primary *JEL* classification dummies; $\widehat{R}_{it}$ was reconstructed at female ratio equal to 1 for women, 0 for men and a paper classified in *JEL* categories D and J. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

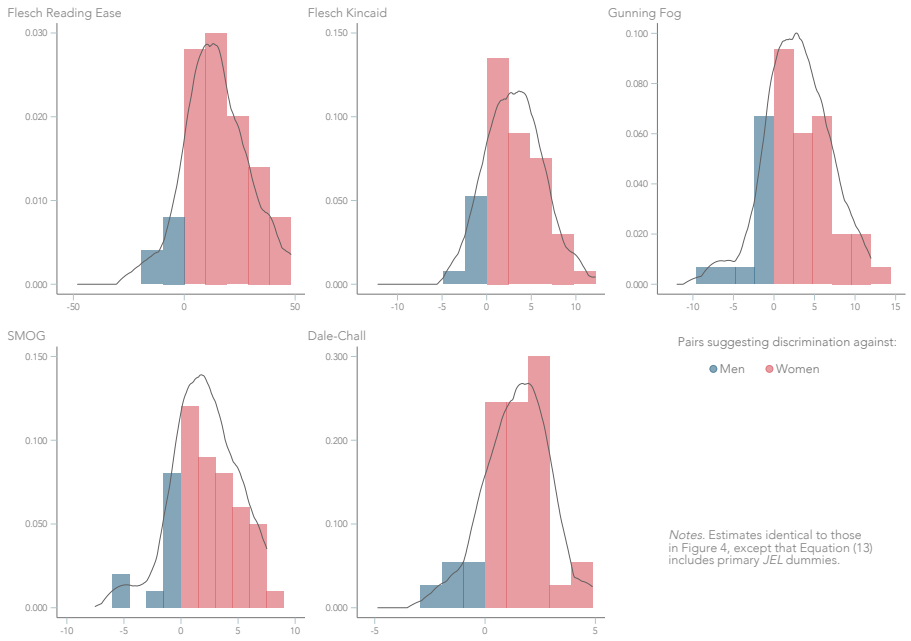


FIGURE F.1: Distributions of $\underline{D}_{ik}$, controlling for *JEL* category

### *F.8 Unadjusted $R_{it}$*

Table F.12 and Figure F.2 replicate the analysis in Section 3.3.2 but Equation (13) does not adjust for the ratio of female authors on a paper. (Thus, $R_{it} = \widehat{R}_{it}$).

TABLE F.12: $\underline{D}_{ik}$, without adjusting for the ratio of female authors

| | Discrimination against women ($\underline{D}_{ik} > 0$) | | | Discrimination against men ($\underline{D}_{ik} < 0$) | | | Mean, all observations | |
|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | $N$ | Mean | S.D. | $N$ | (1) | (2) |
| Flesch Reading Ease | 14.52 | 10.73 | 48 | −8.09 | 8.53 | 25 | 3.76*** | 2.51** |
| | | | | | | | (1.08) | (1.17) |
| Flesch Kincaid | 2.74 | 2.29 | 55 | −2.29 | 2.49 | 23 | 0.81*** | 0.63** |
| | | | | | | | (0.24) | (0.25) |
| Gunning Fog | 3.82 | 2.70 | 45 | −2.39 | 2.84 | 27 | 0.90*** | 0.65** |
| | | | | | | | (0.29) | (0.31) |
| SMOG | 2.77 | 1.99 | 41 | −1.53 | 1.91 | 27 | 0.51** | 0.30 |
| | | | | | | | (0.20) | (0.22) |
| Dale-Chall | 1.37 | 0.90 | 52 | −0.90 | 0.75 | 28 | 0.36*** | 0.24** |
| | | | | | | | (0.11) | (0.12) |

*Notes.* Sample 121 matched pairs (109 and 121 distinct men and women, respectively). Table displays estimates identical to those in Table 7, except that $R_{it}$ is not adjusted for the ratio of female co-authors on a paper. (Thus, $R_{it} = \widehat{R}_{it}$). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

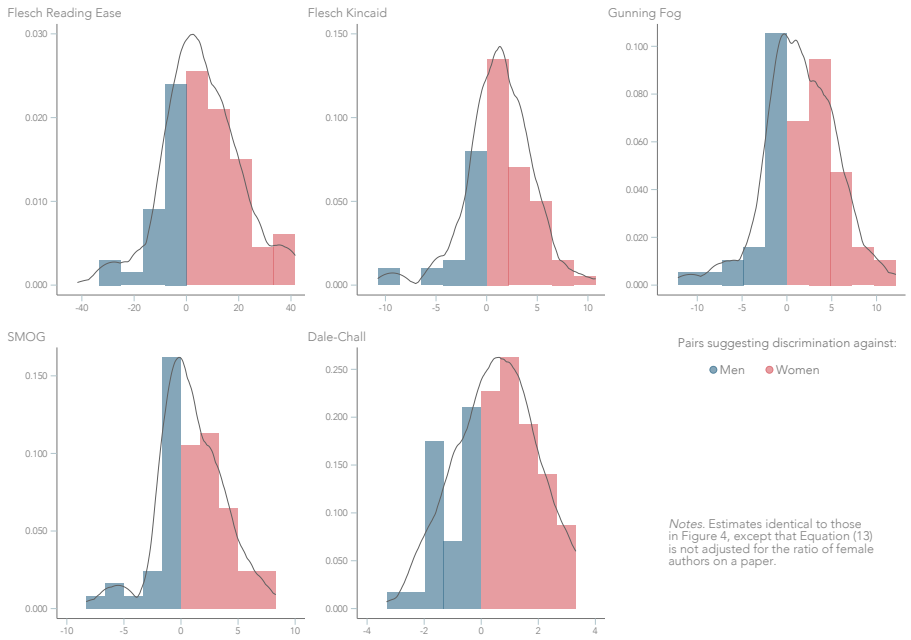


FIGURE F.2: Distributions of $\underline{D}_{ik}$, without adjusting for the ratio of female authors

## G Section 3.4, supplemental output

### *G.1 Table 8, alternative year fixed effects*

Table G.1 and Table G.2 replicate Table 8, replacing acceptance year fixed effects with fixed effects for submission and publication years, respectively.

Table G.1: Table 8 (dependent variable: revision duration), submission year effects

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Female ratio | 2.644** | 4.006*** | 4.035*** | 3.080** | 4.050*** | 4.570** | 4.430** |
| | (1.294) | (1.419) | (1.419) | (1.429) | (1.417) | (1.742) | (1.800) |
| Max. $t_5$ | −0.097*** | −0.099*** | −0.098*** | −0.098*** | −0.097*** | −0.099** | −0.100** |
| | (0.033) | (0.034) | (0.033) | (0.033) | (0.033) | (0.038) | (0.039) |
| No. pages | 0.158*** | 0.156*** | 0.156*** | 0.157*** | 0.156*** | 0.161*** | 0.153*** |
| | (0.020) | (0.020) | (0.020) | (0.020) | (0.020) | (0.025) | (0.028) |
| $N_j$ | 1.013*** | 0.961*** | 0.957*** | 0.991*** | 0.964*** | 0.890*** | 0.802** |
| | (0.244) | (0.240) | (0.243) | (0.240) | (0.241) | (0.285) | (0.305) |
| Order | 0.125** | 0.120* | 0.120* | 0.123** | 0.120* | 0.102 | 0.125 |
| | (0.061) | (0.061) | (0.061) | (0.061) | (0.061) | (0.118) | (0.125) |
| No. citations (asinh) | −0.472*** | −0.486*** | −0.482*** | −0.470*** | −0.491*** | −1.254*** | −1.281*** |
| | (0.164) | (0.165) | (0.164) | (0.164) | (0.165) | (0.283) | (0.273) |
| Flesch Reading Ease | −0.026** | −0.025* | −0.025* | −0.026** | −0.024* | −0.041* | −0.046** |
| | (0.012) | (0.013) | (0.012) | (0.012) | (0.012) | (0.021) | (0.022) |
| Mother | | | −7.160*** | | −10.697*** | −20.405*** | −20.261*** |
| | | | (2.189) | | (3.075) | (2.244) | (2.774) |
| Birth | | | | −3.838 | 5.912 | 16.786*** | 16.677*** |
| | | | | (3.134) | (4.118) | (3.019) | (3.026) |
| Constant | 16.050*** | 16.191*** | 16.191*** | 16.093*** | 16.195*** | 24.408*** | 24.994*** |
| | (1.144) | (1.167) | (1.168) | (1.156) | (1.168) | (2.147) | (2.201) |
| $R^2$ | 0.573 | 0.575 | 0.575 | 0.574 | 0.575 | 0.577 | 0.584 |
| No. observations | 2,620 | 2,605 | 2,620 | 2,620 | 2,620 | 1,278 | 1,278 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Sub. year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Estimates are identical to those in Table 8 except that submission year effects are used instead of acceptance year effects. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE G.2: Table 8 (dependent variable: revision duration), publication year effects

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Female ratio | 5.409*** | 6.786*** | 6.783*** | 5.797*** | 6.802*** | 8.999*** | 9.040*** |
| | (1.578) | (1.929) | (1.931) | (1.847) | (1.928) | (2.398) | (2.462) |
| Max. $t_5$ | −0.137*** | −0.141*** | −0.139*** | −0.138*** | −0.137*** | −0.137*** | −0.143*** |
| | (0.040) | (0.039) | (0.039) | (0.040) | (0.039) | (0.046) | (0.045) |
| No. pages | 0.194*** | 0.194*** | 0.193*** | 0.194*** | 0.193*** | 0.230*** | 0.216*** |
| | (0.026) | (0.026) | (0.026) | (0.026) | (0.026) | (0.039) | (0.041) |
| $N_j$ | 1.148*** | 1.105*** | 1.095*** | 1.129*** | 1.102*** | 1.415*** | 1.293*** |
| | (0.310) | (0.307) | (0.311) | (0.305) | (0.309) | (0.419) | (0.427) |
| Order | 0.212*** | 0.209*** | 0.207*** | 0.210*** | 0.207*** | 0.463*** | 0.452*** |
| | (0.064) | (0.064) | (0.064) | (0.064) | (0.064) | (0.148) | (0.147) |
| No. citations (asinh) | −0.336* | −0.360* | −0.348* | −0.334* | −0.358* | −0.553 | −0.576 |
| | (0.195) | (0.193) | (0.194) | (0.196) | (0.194) | (0.430) | (0.411) |
| Flesch Reading Ease | −0.020 | −0.019 | −0.019 | −0.020 | −0.019 | −0.040 | −0.042 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.026) | (0.027) |
| Mother | | | −7.088** | | −11.096*** | −19.205*** | −19.162*** |
| | | | (3.218) | | (3.254) | (5.214) | (5.509) |
| Birth | | | | −3.429 | 6.689 | 13.651** | 13.336** |
| | | | | (4.020) | (4.095) | (5.238) | (5.480) |
| Constant | 13.756*** | 13.888*** | 13.896*** | 13.791*** | 13.907*** | 16.334*** | 17.275*** |
| | (1.244) | (1.258) | (1.258) | (1.251) | (1.254) | (2.715) | (2.508) |
| $R^2$ | 0.280 | 0.281 | 0.281 | 0.280 | 0.281 | 0.109 | 0.127 |
| No. observations | 2,622 | 2,607 | 2,622 | 2,622 | 2,622 | 1,278 | 1,278 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pub. year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Estimates are identical to those in Table 8 except that publication year effects are used instead of acceptance year effects. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### G.2 *Table 8, alternative thresholds for mother$_j$*

Table G.3 repeats the regression presented in Table 8 column (5), using alternative age thresholds to define motherhood: mother$_j$ equals 1 if paper $j$'s co-authors are all mothers to children younger than three (first column), four (second column), *etc.* Changing this threshold has little effect on female ratio's coefficient. The coefficients on mother$_j$ and birth$_j$ are persistently negative and positive (respectively), although magnitudes and standard errors vary. Remaining coefficients are unaffected.

Table G.3: Table 8 (dependent variable: revision duration), alternative thresholds for mother$_j$

| | Age $< 3$ | Age $< 4$ | Age $< 5$ | Age $< 10$ | Age $< 18$ |
|---|---|---|---|---|---|
| Female ratio | 5.964*** | 6.288*** | 6.922*** | 6.792*** | 6.502*** |
| | (2.101) | (2.079) | (2.085) | (2.121) | (2.220) |
| Mother | −4.957* | −10.511* | −12.299*** | −9.380** | −5.160 |
| | (2.661) | (5.284) | (3.625) | (3.850) | (3.660) |
| Birth | 0.977 | 6.161 | 7.362 | 4.569 | 0.630 |
| | (3.839) | (6.112) | (4.912) | (5.281) | (4.668) |
| Max. $t_5$ | −0.132*** | −0.131*** | −0.130*** | −0.130*** | −0.131*** |
| | (0.038) | (0.038) | (0.038) | (0.039) | (0.039) |
| No. pages | 0.195*** | 0.194*** | 0.194*** | 0.194*** | 0.195*** |
| | (0.028) | (0.028) | (0.028) | (0.028) | (0.027) |
| $N_j$ | 1.137*** | 1.132*** | 1.111*** | 1.110*** | 1.116*** |
| | (0.288) | (0.288) | (0.292) | (0.291) | (0.292) |
| Order | 0.206*** | 0.204*** | 0.203*** | 0.202*** | 0.203*** |
| | (0.064) | (0.064) | (0.064) | (0.064) | (0.064) |
| No. citations (asinh) | −0.398** | −0.403** | −0.420** | −0.414** | −0.407** |
| | (0.197) | (0.197) | (0.196) | (0.196) | (0.197) |
| Flesch Reading Ease | −0.017 | −0.017 | −0.016 | −0.017 | −0.017 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) |
| Constant | 13.911*** | 13.952*** | 14.018*** | 14.030*** | 13.973*** |
| | (1.206) | (1.209) | (1.212) | (1.210) | (1.213) |
| $R^2$ | 0.288 | 0.289 | 0.290 | 0.290 | 0.289 |
| No. observations | 2,622 | 2,622 | 2,622 | 2,622 | 2,622 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review. Coefficients from OLS estimation of Equation (16) at different age thresholds for mother$_j$. In column one, mother$_j$ equals one for papers authored exclusively by women with children younger than three; in column two, the age threshold is four; *etc.* Column three corresponds to results presented in Table 8. Standard errors clustered by year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### G.3 *Quantile regression*

The distribution of review times appears right-skewed (Figure 5). To account for this, I re-estimate gender differences in time spent under review using a quantile regression model. Results are shown in Table G.4: the first panel replicates Table 8, column (5) at the 25th, median and 75th percentiles of review times; the second panel similarly replicates the third column of Table 9.

The coefficient on female ratio is positive and significant across all three percentiles, indicating that the results presented in Section 3.4 are not driven by outliers. Its magnitude is greatest in right-tail of *Econometrica*'s distribution but is similarly sized across all percentiles when estimated using observations from both *Econometrica* and *REStud*.

Table G.4: Revision duration at *Econometrica* and *REStud*, quantile regression

| | *Econometrica* | | | *Econometrica+REStud* | | |
|---|---|---|---|---|---|---|
| | 25th pc. | Median | 75th pc. | 25th pc. | Median | 75th pc. |
| Female ratio | 3.01** | 4.39*** | 6.97*** | 2.36*** | 2.32*** | 2.43** |
| | (1.18) | (1.11) | (1.60) | (0.53) | (0.55) | (1.17) |
| Max. $t_5$ | −0.14*** | −0.16*** | −0.10 | −0.13*** | −0.17*** | −0.13** |
| | (0.03) | (0.04) | (0.07) | (0.02) | (0.02) | (0.06) |
| No. pages | 0.14*** | 0.20*** | 0.26*** | 0.13*** | 0.21*** | 0.26*** |
| | (0.02) | (0.02) | (0.03) | (0.01) | (0.02) | (0.02) |
| $N_j$ | 0.64*** | 0.83*** | 0.79** | 0.47*** | 0.68*** | 0.81*** |
| | (0.19) | (0.24) | (0.34) | (0.08) | (0.14) | (0.20) |
| Order | 0.14*** | 0.19*** | 0.20*** | 0.01 | 0.08** | 0.06 |
| | (0.04) | (0.04) | (0.07) | (0.02) | (0.03) | (0.04) |
| No. citations (asinh) | −0.33*** | −0.46*** | −0.60*** | −0.38*** | −0.34*** | −0.41*** |
| | (0.09) | (0.10) | (0.16) | (0.06) | (0.07) | (0.13) |
| Mother | −5.40 | −11.09 | −10.53** | | | |
| | (34.79) | (11.47) | (4.20) | | | |
| Birth | 5.80 | 7.02 | 13.60*** | | | |
| | (34.76) | (14.65) | (4.15) | | | |
| Constant | 42.27*** | 40.74** | 39.84** | 41.06*** | 37.67*** | 35.58 |
| | (13.01) | (20.20) | (17.29) | (6.91) | (11.63) | (34.55) |
| Pseudo $R^2$ | 0.19 | 0.20 | 0.21 | 0.20 | 0.21 | 0.20 |
| No. observations | 2,623 | 2,623 | 2,623 | 4,435 | 4,435 | 4,435 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | | | |
| Institution effects | ✓ | ✓ | ✓ | | | |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). First panel replicates results shown in Table 8, column (5) across different percentiles of the distribution using quantile regressions; second panel similarly replicates results shown in Table 9, third column. Robust standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

## H Author-level analysis

In this appendix, I analyse readability at the author-level. To disaggregate the data, each article is duplicated $N_j$ times, where $N_j$ is article $j$'s number of co-authors; observation $j_k \in \{1, \ldots, N_j\}$ is assigned article $j$'s $k$th author. I then estimate the dynamic panel model in Equation (H.1):

$$R_{j_{it}} = \beta_0 R_{it-1} + \beta_1 \text{female ratio}_j + \beta_2 \text{female ratio}_j \times \text{male}_i + \boldsymbol{\theta} \mathbf{X}_j + \alpha_i + \varepsilon_{it}. \tag{H.1}$$

$R_{j_{it}}$ is the readability score for article $j$—author $i$'s $t$th top-four publication; $R_{it-1}$ is the corresponding value of author $i$'s $t-1$th top-four paper. Gender enters twice—the binary variable $\text{male}_i$ and $\text{female ratio}_j$—to account for author $i$'s sex and the sex of his co-authors, respectively. $\mathbf{X}_j$ is a vector of observable controls. It includes: editor, journal, year, journal $\times$ year, institution, English fluency dummies and quality controls—citation count (asinh), max. $T_5$ fixed effects (author prominence) and max. $t_5$ (author seniority). I also include $N_j$ to control for author $i$'s proportional contribution to paper $j$.

$\alpha_i$ are author-specific effects and $\varepsilon_{it}$ is an idiosyncratic error. $\alpha_i$ are eliminated by first-differencing. For each time period, endogeneity in the lagged dependant variable is instrumented with up to five earlier lags (Arellano and Bover, 1995; Blundell and Bond, 1998).[39] To account for duplicate articles, the regression is weighted by $1/N_j$.[40] Standard errors are adjusted for two-way clustering on editor and author.

Table H.1 displays results. Rows one and two present contemporaneous marginal effects on co-authoring with women for female ($\beta_1$) and male ($\beta_1 + \beta_2$) authors, respectively. Both estimates are positive—everyone writes more clearly when collaborating with women—although statistically significant only for female authors. Marginal effects for women are twice as large as those shown in Table 2; they suggest women write 2–6 percent better than men.[41]

Men and women co-authoring together experience an identical rise (or fall) in readability, so the effect for one should mirror the other. Yet, Table H.1 suggests they don't. While the interaction terms ($\beta_2$) are insignificant, the presence of a difference could indicate an increasing, convex relationship between female ratio and readability.[42] Thus, men's smaller effect potentially reflects their disproportionate tendency to co-author exclusively with other men—*i.e.*, precisely where the marginal impact of an additional woman is low.[43]

Coefficients on the lagged dependant variables are small, suggesting readability is mostly determined contemporaneously. Nevertheless, their uniform positivity and significance indicate modest persistence.

Table H.1's second panel report test statistics of model fit. The first two rows test for serial correlation; they indicate no model misspecification. $p$-values on the overall Hansen test statistic hover between 0.35–0.72. Additional tests (available on request) suggest results are not sensitive to including the full set of (non-collapsed) instruments or to reductions in the number of instruments.

[39] Results are robust to not collapsing instruments and instrumenting only with one up to all earlier lags (available on request).

[40] Assigning equal weight to all observations results in quantitatively and qualitatively similar results (see Hengel, 2016, pp. 44–45).

[41] Quotient of $\beta_1$ divided by the total effect for men co-authoring with no women (female ratio of zero) estimated at other co-variates' observed values.

[42] I re-estimated Equation (H.1) replacing $\text{female ratio}_j \times \text{male}_i$ with $\text{female ratio}_j^2$. The results—which are available on request—tentatively support this conclusion.

[43] On average, the female ratio for men is 0.04 (0.05 excluding solo-authored papers). When excluding articles written entirely by men, their average ratio is still only 0.39. By default, women always author with at least one woman—themselves; the average female ratio of their papers is 0.6 (0.46 and 0.53 excluding articles written entirely by women and solo-authored papers, respectively).

TABLE H.1: Gender differences in readability, author-level analysis

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 2.31** | 0.27 | 0.52** | 0.39** | 0.26*** |
| | (0.93) | (0.20) | (0.24) | (0.17) | (0.09) |
| Female ratio for men ($\beta_1 + \beta_2$) | 0.44 | 0.08 | 0.12 | 0.06 | 0.10 |
| | (1.26) | (0.26) | (0.31) | (0.22) | (0.11) |
| Female ratio×male ($\beta_2$) | −1.87 | −0.19 | −0.40 | −0.33 | −0.16 |
| | (1.50) | (0.31) | (0.37) | (0.27) | (0.14) |
| Lagged score ($\beta_0$) | 0.04* | 0.04* | 0.03 | 0.03 | 0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.35 | 0.72 | 0.34 | 0.71 | 0.34 |
| *$z$-test for no serial correlation* | | | | | |
| Order 1 | −18.53 | −13.61 | −14.97 | −17.57 | −18.76 |
| Order 2 | 0.72 | −0.27 | 0.28 | 0.52 | 0.28 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Figures from first-differenced, IV estimation of Equation (H.1) (Arellano and Bover, 1995; Blundell and Bond, 1998) where instruments have been collapsed to create 186 total instrument: one instrument for each variable and lag distance (see Footnote 39). Female ratio (women): contemporaneous marginal effect of a paper's female co-author ratio for female authors ($\beta_1$); female ratio (men): analogous effect for male authors ($\beta_1 + \beta_2$). $z$-statistics for first- and second-order autocorrelation in the first-differenced errors (Arellano and Bond, 1991); null hypothesis no autocorrelation. Quality controls denoted by ✓[2] include citation count (asinh), max. $T_5$ (author prominence) and max. $t_5$ (author seniority); to reduce the number of instruments, institutions are entered as levels. Standard errors clustered on author (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

# I Alternative program for calculating readability scores

In this section, I replicate Table 2, Table 4, Table H.1 and Table 6 using readability scores generated by the R `readability` package, an alternative program for calculating Flesch-Kincaid, Gunning Fog and SMOG readability scores.[44] Replications for other tables and figures presented in the paper are not shown, but will be made available on request.

`Textatistic` and `readability` employ different strategies to adapt the scores to automated calculation—*e.g.*, `readability` counts semi-colons and dashes as sentence-ending terminations; `Textatistic` does not.[45] Results appear robust to these (and other) small discrepancies: coefficients are similar to those presented in the body of the paper; standard errors are usually smaller.

TABLE I.1: Table 2, alternative program for calculating readability

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch-Kincaid | 0.19* | 0.19* | 0.18* | 0.20* | 0.22* | 0.25* | 0.27** | 0.28* |
| | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.13) | (0.12) | (0.14) |
| Gunning Fog | 0.33*** | 0.33*** | 0.34*** | 0.36*** | 0.38*** | 0.38*** | 0.36*** | 0.33** |
| | (0.11) | (0.11) | (0.12) | (0.11) | (0.11) | (0.13) | (0.13) | (0.16) |
| SMOG | 0.21** | 0.21** | 0.22** | 0.23*** | 0.25*** | 0.24** | 0.24** | 0.21* |
| | (0.09) | (0.09) | (0.09) | (0.09) | (0.08) | (0.10) | (0.09) | (0.12) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 9,117 articles in (1)–(5); 5,211 articles in (6) and (7); 5,774 articles—including 563 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Figures are identical to those in Table 2, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

[44] The program does not calculate the Flesch Reading Ease or Dale-Chall scores.

[45] Readability scores were originally developed to be calculated by hand. Automating their calculation requires slightly adapting the algorithms. For example, all five scores define sentences as grammatically independent units of thoughts—*e.g.*, two independent clauses connected by a dash or semi-colon count as two separate sentences. Unfortunately, semi-colons and dashes are frequently used in other ways and it is difficult to programmatically distinguish between contexts.

TABLE I.2: Table 4, alternative program for calculating readability

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch-Kincaid | 0.51*** | 0.48** | 0.87*** | 0.40** | 0.40** |
| | (0.17) | (0.22) | (0.29) | (0.19) | (0.19) |
| Gunning Fog | 0.51*** | 0.63** | 1.00*** | 0.38** | 0.38* |
| | (0.18) | (0.25) | (0.28) | (0.19) | (0.19) |
| SMOG | 0.37** | 0.42*** | 0.69*** | 0.27** | 0.27** |
| | (0.14) | (0.16) | (0.19) | (0.13) | (0.14) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,709 NBER working papers; 1,707 published articles. Figures are identical to those in Table 4, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE I.3: Table 6, alternative program for calculating readability

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch-Kincaid | 0.12 | 0.18 | 0.91*** | 0.57 | 0.52 | 0.22 |
| | (0.14) | (0.22) | (0.27) | (0.43) | (0.37) | (0.16) |
| Gunning Fog | 0.31** | 0.27 | 1.19*** | 0.75 | 0.71 | 0.44** |
| | (0.16) | (0.25) | (0.37) | (0.51) | (0.47) | (0.19) |
| SMOG | 0.20* | 0.18 | 0.77*** | 0.50 | 0.50 | 0.31** |
| | (0.12) | (0.18) | (0.25) | (0.37) | (0.33) | (0.13) |
| No. observations | 6,875 | 2,826 | 1,675 | 1,906 | 2,773 | 12,006 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Figures are identical to those in Table 6, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE I.4: Table H.1, alternative program for calculating readability

| | Flesch-Kincaid | Gunning Fog | SMOG |
|---|---|---|---|
| Female ratio for women ($\beta_1$) | 0.30 | 0.49** | 0.31* |
| | (0.21) | (0.25) | (0.18) |
| Female ratio for men ($\beta_1 + \beta_2$) | 0.13 | 0.27 | 0.17 |
| | (0.27) | (0.31) | (0.22) |
| Female ratio×male ($\beta_2$) | −0.17 | −0.22 | −0.14 |
| | (0.33) | (0.38) | (0.27) |
| Lagged score ($\beta_0$) | 0.05** | 0.04* | 0.04* |
| | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.47 | 0.30 | 0.29 |
| ***z-test for no serial correlation*** | | | |
| Order 1 | −12.66 | −13.49 | −16.97 |
| Order 2 | 0.42 | 0.68 | 0.51 |
| Editor effects | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Figures are identical to those in Table H.1, except readability scores were calculated using the R `readability` program. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

## J Alternative proxies for article gender

The following sections replicate Table 2, Table 4, Table 6, Table 8, Table 9 and Table H.1 using alternative proxies for article gender.[46] Replications using additional proxies are available on request (subject to feasibility).

- In Appendix J.1, the sample is restricted to solo-authored papers.
- In Appendix J.2, article gender is represented by a binary variable equal to one if the author with the (weakly) greatest number of top-five papers at the time of publication is female; mixed-gendered articles without a senior female co-author are excluded.
- In Appendix J.3, article gender is represented by a binary variable equal to one if at least half of all authors are female; mixed-gendered articles below this threshold are excluded.
- In Appendix J.4, a paper is considered "female" if at least one author is female.
- In Appendix J.5, papers authored entirely by women are compared to papers authored entirely by men. Co-authored mixed-sex articles are excluded.
- In Appendix J.6, the sample includes only articles written by non-experienced authors (defined as having two or fewer previous top-five articles) and article gender is represented by a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published.[47]

The estimation strategy in Section 3.3 relies on within-author differences in readability scores at two specific $t_4$ ($t_4 = 1$ and $t_4 = 3$). Because only a small number of women have majority and exclusively female-authored papers for both $t_4$, I reproduce instead results from Table 6 using the alternative proxies for article gender.

In general, standard errors are smaller and coefficients larger in Appendix J.2, Appendix J.3 and Appendix J.4; the reverse is usually—but not always—true for Appendix J.5 (which includes a much smaller number of female-authored papers); coefficients are similar in Appendix J.1, but standard errors are also larger (again, possibly due to small sample sizes).

[46] In order to generate within-author variation, mixed-sex co-authored articles that do not satisfy the relevant "female" definition are included—but classified as male—when estimating Table H.1. Otherwise, these observations are excluded.

[47] Note that Table 6 and Table H.1 cannot be estimated given the sample restriction on author experience. Additionally, the results in tables Table J.33 and Table J.34 contain very few female observations with *experienced* senior female authors and are therefore similar to the results shown in Table 8 and Table 9, respectively. See Figure 7 for a breakdown of review times between junior and senior women where "junior" is defined as having at most only one previous top-five publication.

### *J.1 Solo-authored*

TABLE J.1: Table 2, solo-authored papers

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.49 | 0.41 | 0.34 | 0.51 | 0.79 | 0.52 | 0.61 | 0.75 |
| | (0.57) | (0.56) | (0.57) | (0.60) | (0.60) | (0.66) | (0.71) | (0.90) |
| Flesch-Kincaid | 0.11 | 0.09 | 0.08 | 0.13 | 0.16 | 0.21 | 0.24 | 0.21 |
| | (0.13) | (0.13) | (0.13) | (0.14) | (0.14) | (0.15) | (0.15) | (0.18) |
| Gunning Fog | 0.22 | 0.21 | 0.21 | 0.27* | 0.31* | 0.41** | 0.41** | 0.35* |
| | (0.14) | (0.14) | (0.14) | (0.15) | (0.16) | (0.17) | (0.17) | (0.20) |
| SMOG | 0.15 | 0.14 | 0.14 | 0.17 | 0.21* | 0.25** | 0.25** | 0.22 |
| | (0.10) | (0.10) | (0.10) | (0.11) | (0.11) | (0.12) | (0.12) | (0.15) |
| Dale-Chall | 0.06 | 0.06 | 0.05 | 0.06 | 0.08 | 0.13** | 0.13* | 0.15** |
| | (0.05) | (0.05) | (0.05) | (0.06) | (0.06) | (0.06) | (0.07) | (0.07) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 4,014 articles in (1)–(5); 1,540 articles in (6) and (7); 1,666 articles—including 126 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored papers are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.2: Table 4, solo-authored papers

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.72 | 1.37 | 2.88 | 1.50 | 1.50 |
| | (1.67) | (2.08) | (2.38) | (1.19) | (1.29) |
| Flesch-Kincaid | 0.64 | 0.31 | 0.92* | 0.61* | 0.61* |
| | (0.50) | (0.43) | (0.54) | (0.33) | (0.36) |
| Gunning Fog | 0.71 | 0.37 | 1.03* | 0.66* | 0.66 |
| | (0.56) | (0.49) | (0.61) | (0.38) | (0.41) |
| SMOG | 0.41 | 0.27 | 0.63 | 0.36 | 0.36 |
| | (0.33) | (0.34) | (0.39) | (0.25) | (0.27) |
| Dale-Chall | 0.25 | 0.26* | 0.47*** | 0.21 | 0.21 |
| | (0.16) | (0.15) | (0.16) | (0.14) | (0.15) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 345 NBER working papers; 344 published articles (37 female-authored). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored papers are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.3: Table 6, solo-authored papers

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|
| Flesch Reading Ease | −0.31 | 0.85 | 2.64 | 9.40* | 0.47 |
| | (1.01) | (1.54) | (2.60) | (4.83) | (0.84) |
| Flesch-Kincaid | −0.01 | −0.06 | 0.45 | 2.32*** | 0.05 |
| | (0.24) | (0.36) | (0.74) | (0.54) | (0.19) |
| Gunning Fog | 0.10 | −0.23 | 0.79 | 2.90*** | 0.15 |
| | (0.27) | (0.44) | (0.84) | (0.79) | (0.22) |
| SMOG | 0.06 | −0.18 | 0.63 | 1.99*** | 0.11 |
| | (0.18) | (0.30) | (0.59) | (0.76) | (0.15) |
| Dale-Chall | −0.09 | 0.28** | 0.29 | 0.99** | 0.01 |
| | (0.10) | (0.13) | (0.21) | (0.45) | (0.07) |
| No. observations | 2,025 | 758 | 772 | 459 | 4,013 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored papers are excluded.) Due to small sample sizes, final column estimates are clustered on author, only. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.4: Table 8 (dependent variable: revision duration), solo-authored papers

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Solo female | 6.375** | 9.417** | 9.200** | 7.218** | 9.291** | 11.526* | 12.658** |
| | (2.886) | (3.989) | (3.951) | (3.498) | (3.952) | (5.661) | (5.893) |
| Max. $t_5$ | −0.314*** | −0.309*** | −0.311*** | −0.315*** | −0.307*** | −0.286*** | −0.274** |
| | (0.076) | (0.076) | (0.076) | (0.077) | (0.076) | (0.098) | (0.105) |
| No. pages | 0.196*** | 0.193*** | 0.191*** | 0.195*** | 0.191*** | 0.265*** | 0.265*** |
| | (0.042) | (0.042) | (0.042) | (0.042) | (0.041) | (0.064) | (0.068) |
| Order | 0.168* | 0.168* | 0.163* | 0.166* | 0.164* | 0.545* | 0.428 |
| | (0.084) | (0.084) | (0.083) | (0.084) | (0.083) | (0.306) | (0.318) |
| No. citations (asinh) | −0.735*** | −0.775*** | −0.741*** | −0.726*** | −0.760*** | −1.306** | −1.209** |
| | (0.180) | (0.176) | (0.178) | (0.182) | (0.178) | (0.594) | (0.554) |
| Flesch Reading Ease | −0.011 | −0.010 | −0.009 | −0.011 | −0.007 | −0.006 | 0.006 |
| | (0.022) | (0.022) | (0.022) | (0.022) | (0.022) | (0.062) | (0.067) |
| Mother | | | −8.880* | | −13.676*** | −28.084*** | −28.926*** |
| | | | (4.604) | | (5.026) | (8.465) | (8.537) |
| Birth | | | | −4.407 | 7.496 | 22.305*** | 24.844*** |
| | | | | (5.011) | (4.668) | (6.871) | (7.477) |
| Constant | 15.864*** | 16.031*** | 15.925*** | 15.871*** | 15.945*** | 19.926*** | 19.486*** |
| | (1.800) | (1.761) | (1.793) | (1.806) | (1.780) | (5.598) | (5.392) |
| $R^2$ | 0.336 | 0.341 | 0.340 | 0.337 | 0.341 | 0.211 | 0.242 |
| No. observations | 1,223 | 1,209 | 1,223 | 1,223 | 1,223 | 417 | 415 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored papers are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.5: Table 9 (dependent variable: revision duration), solo-authored papers

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Solo female | 6.34** | 2.15 | 3.95** | 10.43** | 4.25 | 6.55*** |
| | (2.90) | (1.68) | (1.54) | (4.72) | (2.64) | (2.23) |
| Max. $t_5$ | −0.29*** | −0.38*** | −0.33*** | −0.31*** | −0.39*** | −0.35*** |
| | (0.08) | (0.09) | (0.06) | (0.10) | (0.12) | (0.07) |
| No. pages | 0.20*** | 0.24** | 0.21*** | 0.25*** | 0.06 | 0.20*** |
| | (0.04) | (0.10) | (0.04) | (0.07) | (0.16) | (0.05) |
| Order | 0.15* | −0.05 | 0.03 | 0.41 | 0.05 | 0.24 |
| | (0.08) | (0.09) | (0.06) | (0.32) | (0.25) | (0.21) |
| No. citations (asinh) | −0.60*** | −0.42 | −0.51*** | −0.90* | −1.41** | −1.23** |
| | (0.19) | (0.28) | (0.17) | (0.48) | (0.64) | (0.49) |
| Constant | 14.91*** | 20.28*** | 17.30*** | 19.11*** | 32.55*** | 24.83*** |
| | (1.36) | (2.19) | (1.09) | (4.29) | (4.81) | (3.03) |
| $R^2$ | 0.33 | 0.34 | 0.36 | 0.21 | 0.25 | 0.22 |
| No. observations | 1,223 | 850 | 2,073 | 415 | 374 | 790 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal × Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored papers are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.6: Table H.1, solo-authored papers

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 1.96 | 0.17 | 0.41 | 0.35 | 0.30** |
| | (1.38) | (0.30) | (0.37) | (0.27) | (0.12) |
| Lagged score ($\beta_0$) | 0.04* | 0.05** | 0.03 | 0.03 | 0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.32 | 0.68 | 0.04 | 0.65 | 0.47 |
| *$z$-test for no serial correlation* | | | | | |
| Order 1 | −18.47 | −13.69 | −14.90 | −17.52 | −18.58 |
| Order 2 | 0.70 | −0.17 | 0.27 | 0.51 | 0.30 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,132 observations (2,812 authors, 121 female). Estimates identical to those in Table H.1, except that female ratio has been replaced with a dummy variable equal to 1 if the paper is solo-authored by a woman and 0 if it is solo-authored by a man. (Co-authored mixed-sex papers are included and classified as male (see Footnote 46).) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

## *J.2 Senior female author*

TABLE J.7: Table 2, senior female author

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.84* | 0.79 | 0.73 | 0.82 | 1.04** | 0.83 | 0.82 | 0.82 |
| | (0.50) | (0.49) | (0.51) | (0.52) | (0.52) | (0.55) | (0.56) | (0.64) |
| Flesch-Kincaid | 0.17* | 0.16 | 0.16 | 0.18* | 0.20* | 0.24** | 0.25** | 0.21 |
| | (0.10) | (0.10) | (0.10) | (0.11) | (0.11) | (0.11) | (0.11) | (0.13) |
| Gunning Fog | 0.29** | 0.29*** | 0.29** | 0.33*** | 0.35*** | 0.41*** | 0.38*** | 0.33** |
| | (0.11) | (0.11) | (0.11) | (0.12) | (0.12) | (0.12) | (0.12) | (0.14) |
| SMOG | 0.19** | 0.19** | 0.19** | 0.21** | 0.23** | 0.26** | 0.23** | 0.20** |
| | (0.08) | (0.08) | (0.08) | (0.09) | (0.09) | (0.09) | (0.09) | (0.10) |
| Dale-Chall | 0.06 | 0.06 | 0.06 | 0.06 | 0.08 | 0.11* | 0.10* | 0.12** |
| | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.06) | (0.05) | (0.06) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 8,917 articles in (1)–(5); 5,045 articles in (6) and (7); 5,557 articles—including 512 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.8: Table 4, senior female author

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.62*** | 1.81** | 3.12*** | 1.31** | 1.31** |
| | (0.62) | (0.78) | (0.95) | (0.64) | (0.65) |
| Flesch-Kincaid | 0.53*** | 0.29** | 0.75*** | 0.46** | 0.46** |
| | (0.14) | (0.13) | (0.22) | (0.19) | (0.19) |
| Gunning Fog | 0.57*** | 0.41*** | 0.89*** | 0.48** | 0.48** |
| | (0.14) | (0.15) | (0.24) | (0.20) | (0.21) |
| SMOG | 0.34*** | 0.32*** | 0.59*** | 0.27** | 0.27** |
| | (0.10) | (0.12) | (0.16) | (0.12) | (0.13) |
| Dale-Chall | 0.10 | 0.24** | 0.30*** | 0.06 | 0.06 |
| | (0.07) | (0.10) | (0.10) | (0.06) | (0.06) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,655 NBER working papers; 1,653 published articles (119 female-authored). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.9: Table 6, senior female author

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch Reading Ease | −0.01 | 1.22 | 2.93** | 2.82* | 1.76 | 1.57** |
| | (0.80) | (0.82) | (1.18) | (1.47) | (1.63) | (0.74) |
| Flesch-Kincaid | −0.03 | 0.19 | 0.58** | 0.53* | 0.32 | 0.23 |
| | (0.17) | (0.15) | (0.27) | (0.32) | (0.35) | (0.15) |
| Gunning Fog | 0.07 | 0.39** | 0.85** | 0.88** | 0.42 | 0.49*** |
| | (0.20) | (0.19) | (0.37) | (0.35) | (0.40) | (0.19) |
| SMOG | 0.04 | 0.25* | 0.56** | 0.70*** | 0.28 | 0.35*** |
| | (0.14) | (0.14) | (0.26) | (0.26) | (0.30) | (0.13) |
| Dale-Chall | 0.00 | 0.13 | 0.20 | 0.18 | 0.25 | 0.14* |
| | (0.07) | (0.09) | (0.15) | (0.15) | (0.19) | (0.08) |
| No. observations | 6,539 | 2,749 | 1,646 | 1,871 | 2,766 | 11,472 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.10: Table 8 (dependent variable: revision duration), senior female author

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Senior female | 4.978** | 6.790** | 6.786** | 5.496** | 6.805** | 8.310** | 8.368** |
| | (2.314) | (2.888) | (2.869) | (2.690) | (2.871) | (3.585) | (3.503) |
| Max. $t_5$ | −0.129*** | −0.129*** | −0.127*** | −0.129*** | −0.126*** | −0.124** | −0.129*** |
| | (0.039) | (0.039) | (0.039) | (0.039) | (0.039) | (0.045) | (0.047) |
| No. pages | 0.194*** | 0.193*** | 0.192*** | 0.194*** | 0.192*** | 0.233*** | 0.219*** |
| | (0.028) | (0.028) | (0.028) | (0.028) | (0.028) | (0.041) | (0.044) |
| $N_j$ | 1.284*** | 1.275*** | 1.264*** | 1.274*** | 1.272*** | 1.663*** | 1.521*** |
| | (0.310) | (0.305) | (0.310) | (0.305) | (0.307) | (0.413) | (0.438) |
| Order | 0.211*** | 0.209*** | 0.207*** | 0.210*** | 0.207*** | 0.480*** | 0.468*** |
| | (0.065) | (0.065) | (0.065) | (0.065) | (0.065) | (0.149) | (0.145) |
| No. citations (asinh) | −0.368* | −0.386* | −0.375* | −0.364* | −0.386* | −0.553 | −0.545 |
| | (0.199) | (0.198) | (0.199) | (0.199) | (0.199) | (0.406) | (0.393) |
| Flesch Reading Ease | −0.018 | −0.016 | −0.016 | −0.017 | −0.016 | −0.036 | −0.038 |
| | (0.014) | (0.015) | (0.014) | (0.014) | (0.014) | (0.029) | (0.030) |
| Mother | | | −7.887** | | −12.042*** | −20.959*** | −21.075*** |
| | | | (3.792) | | (3.944) | (6.580) | (7.057) |
| Birth | | | | −3.837 | 6.910 | 15.510** | 15.231** |
| | | | | (4.795) | (4.852) | (6.404) | (6.658) |
| Constant | 13.645*** | 13.709*** | 13.725*** | 13.663*** | 13.734*** | 15.798*** | 16.628*** |
| | (1.241) | (1.255) | (1.250) | (1.245) | (1.248) | (2.541) | (2.381) |
| $R^2$ | 0.287 | 0.290 | 0.288 | 0.287 | 0.289 | 0.124 | 0.143 |
| No. observations | 2,588 | 2,573 | 2,588 | 2,588 | 2,588 | 1,253 | 1,253 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.11: Table 9 (dependent variable: revision duration), senior female author

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Senior female | 4.87** | 1.58 | 3.08*** | 6.41** | 3.08 | 4.35*** |
| | (2.34) | (1.54) | (1.11) | (3.04) | (2.08) | (1.48) |
| Max. $t_5$ | −0.12*** | −0.10 | −0.12*** | −0.12*** | −0.05 | −0.09** |
| | (0.04) | (0.06) | (0.04) | (0.04) | (0.07) | (0.04) |
| No. pages | 0.20*** | 0.13** | 0.18*** | 0.21*** | 0.05 | 0.17*** |
| | (0.03) | (0.06) | (0.03) | (0.04) | (0.07) | (0.04) |
| $N_j$ | 1.29*** | −0.07 | 0.80*** | 1.46*** | 0.27 | 1.11*** |
| | (0.31) | (0.51) | (0.28) | (0.45) | (0.68) | (0.37) |
| Order | 0.20*** | −0.06 | 0.09* | 0.47*** | 0.06 | 0.23** |
| | (0.06) | (0.08) | (0.05) | (0.13) | (0.15) | (0.11) |
| No. citations (asinh) | −0.27 | −0.60** | −0.39** | −0.32 | −1.20** | −0.78** |
| | (0.19) | (0.23) | (0.16) | (0.39) | (0.45) | (0.31) |
| Constant | 12.51*** | 23.51*** | 16.67*** | 14.40*** | 30.52*** | 21.27*** |
| | (1.20) | (1.85) | (0.95) | (2.11) | (3.15) | (1.57) |
| $R^2$ | 0.28 | 0.27 | 0.29 | 0.13 | 0.14 | 0.13 |
| No. observations | 2,588 | 1,778 | 4,366 | 1,253 | 1,040 | 2,294 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.12: Table H.1, senior female author

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 2.02** | 0.33* | 0.58** | 0.40** | 0.15* |
| | (0.94) | (0.20) | (0.23) | (0.16) | (0.09) |
| Female ratio for men ($\beta_1 + \beta_2$) | 2.48 | 0.44 | 0.50 | 0.34 | 0.03 |
| | (1.76) | (0.39) | (0.46) | (0.32) | (0.14) |
| Female ratio×male ($\beta_2$) | 0.46 | 0.10 | −0.08 | −0.07 | −0.12 |
| | (1.97) | (0.43) | (0.51) | (0.36) | (0.17) |
| Lagged score ($\beta_0$) | 0.04** | 0.05** | 0.03 | 0.03 | 0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.35 | 0.72 | 0.46 | 0.71 | 0.05 |
| *z-test for no serial correlation* | | | | | |
| Order 1 | −18.53 | −13.74 | −14.89 | −17.56 | −18.77 |
| Order 2 | 0.73 | −0.17 | 0.21 | 0.52 | 0.25 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Estimates identical to those in Table H.1, except that female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. Otherwise, it is 0. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *J.3 Majority female-authored*

TABLE J.13: Table 2, majority female-authored

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.93*** | 0.91*** | 0.89*** | 0.93*** | 1.10*** | 0.91** | 0.95** | 1.13** |
| | (0.32) | (0.32) | (0.33) | (0.33) | (0.33) | (0.34) | (0.35) | (0.45) |
| Flesch-Kincaid | 0.16** | 0.16* | 0.16* | 0.17** | 0.19** | 0.23** | 0.26*** | 0.26** |
| | (0.08) | (0.08) | (0.08) | (0.08) | (0.09) | (0.09) | (0.09) | (0.10) |
| Gunning Fog | 0.27*** | 0.27*** | 0.27*** | 0.28*** | 0.30*** | 0.33*** | 0.33*** | 0.34*** |
| | (0.09) | (0.10) | (0.10) | (0.10) | (0.09) | (0.11) | (0.10) | (0.12) |
| SMOG | 0.17** | 0.17** | 0.17** | 0.18*** | 0.20*** | 0.20*** | 0.20*** | 0.23** |
| | (0.07) | (0.07) | (0.07) | (0.07) | (0.06) | (0.07) | (0.07) | (0.09) |
| Dale-Chall | 0.09** | 0.09** | 0.08** | 0.09** | 0.09** | 0.09** | 0.09** | 0.11** |
| | (0.03) | (0.03) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.05) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 8,800 articles in (1)–(5); 4,913 articles in (6) and (7); 5,403 articles—including 490 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.14: Table 4, majority female-authored

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.34*** | 1.30** | 2.43*** | 1.12*** | 1.12*** |
| | (0.43) | (0.66) | (0.72) | (0.41) | (0.42) |
| Flesch-Kincaid | 0.44*** | 0.16 | 0.56*** | 0.40*** | 0.40*** |
| | (0.14) | (0.16) | (0.16) | (0.12) | (0.12) |
| Gunning Fog | 0.46*** | 0.21 | 0.62*** | 0.41*** | 0.41*** |
| | (0.16) | (0.18) | (0.18) | (0.13) | (0.13) |
| SMOG | 0.28*** | 0.14 | 0.40*** | 0.25*** | 0.25*** |
| | (0.11) | (0.11) | (0.12) | (0.09) | (0.09) |
| Dale-Chall | 0.14*** | 0.16*** | 0.27*** | 0.12*** | 0.12*** |
| | (0.03) | (0.05) | (0.06) | (0.03) | (0.03) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,567 NBER working papers; 1,565 published articles (235 female-authored). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.15: Table 6, majority female-authored

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–5 | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch Reading Ease | 0.58 | 1.42** | 3.63*** | 1.97 | 2.36* | 2.04*** |
| | (0.51) | (0.59) | (1.03) | (1.41) | (1.27) | (0.65) |
| Flesch-Kincaid | 0.08 | 0.16 | 0.71*** | 0.56** | 0.42 | 0.32** |
| | (0.12) | (0.14) | (0.19) | (0.25) | (0.30) | (0.14) |
| Gunning Fog | 0.20 | 0.26 | 0.95*** | 0.69** | 0.53 | 0.45*** |
| | (0.13) | (0.18) | (0.24) | (0.28) | (0.34) | (0.17) |
| SMOG | 0.13 | 0.19 | 0.60*** | 0.46** | 0.36 | 0.31*** |
| | (0.09) | (0.13) | (0.18) | (0.22) | (0.24) | (0.12) |
| Dale-Chall | 0.08* | 0.07 | 0.22** | 0.18 | 0.27 | 0.15** |
| | (0.05) | (0.06) | (0.11) | (0.11) | (0.17) | (0.06) |
| No. observations | 6,402 | 2,679 | 1,557 | 1,777 | 2,577 | 9,587 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.16: Table 8 (dependent variable: revision duration), majority female-authored

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Majority female | 3.347*** | 3.799*** | 3.786*** | 3.414*** | 3.802*** | 5.849*** | 5.826*** |
| | (1.151) | (1.291) | (1.288) | (1.268) | (1.286) | (1.677) | (1.735) |
| Max. $t_5$ | −0.168*** | −0.172*** | −0.170*** | −0.169*** | −0.168*** | −0.174*** | −0.170*** |
| | (0.039) | (0.039) | (0.039) | (0.039) | (0.039) | (0.047) | (0.047) |
| No. pages | 0.201*** | 0.201*** | 0.200*** | 0.201*** | 0.200*** | 0.244*** | 0.229*** |
| | (0.030) | (0.030) | (0.031) | (0.030) | (0.031) | (0.044) | (0.046) |
| $N_j$ | 1.200*** | 1.186*** | 1.171*** | 1.194*** | 1.179*** | 1.590*** | 1.398*** |
| | (0.324) | (0.317) | (0.321) | (0.315) | (0.318) | (0.421) | (0.434) |
| Order | 0.197*** | 0.195*** | 0.193*** | 0.196*** | 0.193*** | 0.446*** | 0.441*** |
| | (0.061) | (0.062) | (0.062) | (0.062) | (0.062) | (0.144) | (0.141) |
| No. citations (asinh) | −0.412** | −0.430** | −0.418** | −0.411** | −0.429** | −0.709* | −0.666 |
| | (0.200) | (0.199) | (0.199) | (0.200) | (0.199) | (0.415) | (0.405) |
| Flesch Reading Ease | −0.014 | −0.013 | −0.013 | −0.014 | −0.013 | −0.032 | −0.036 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.029) | (0.029) |
| Mother | | | −4.391 | | −8.632*** | −16.519*** | −16.708*** |
| | | | (2.935) | | (2.837) | (4.285) | (4.990) |
| Birth | | | | −1.164 | 7.063 | 14.384*** | 14.815** |
| | | | | (4.118) | (4.443) | (5.173) | (5.503) |
| Constant | 13.824*** | 13.904*** | 13.919*** | 13.837*** | 13.931*** | 16.450*** | 17.214*** |
| | (1.239) | (1.257) | (1.255) | (1.245) | (1.252) | (2.623) | (2.404) |
| $R^2$ | 0.287 | 0.288 | 0.287 | 0.287 | 0.288 | 0.120 | 0.139 |
| No. observations | 2,543 | 2,528 | 2,543 | 2,543 | 2,543 | 1,211 | 1,211 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.17: Table 9 (dependent variable: revision duration), majority female-authored

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Majority female | 3.22*** | 0.89 | 1.98*** | 5.14*** | 2.14** | 3.51*** |
| | (1.16) | (0.83) | (0.74) | (1.56) | (0.99) | (0.89) |
| Max. $t_5$ | −0.16*** | −0.10 | −0.14*** | −0.16*** | −0.05 | −0.11** |
| | (0.04) | (0.06) | (0.04) | (0.04) | (0.07) | (0.04) |
| No. pages | 0.21*** | 0.13* | 0.18*** | 0.23*** | 0.04 | 0.17*** |
| | (0.03) | (0.07) | (0.03) | (0.04) | (0.07) | (0.04) |
| $N_j$ | 1.22*** | −0.18 | 0.72** | 1.35*** | 0.20 | 1.04** |
| | (0.33) | (0.48) | (0.29) | (0.46) | (0.67) | (0.38) |
| Order | 0.18*** | −0.07 | 0.08 | 0.44*** | 0.05 | 0.21* |
| | (0.06) | (0.08) | (0.05) | (0.13) | (0.16) | (0.12) |
| No. citations (asinh) | −0.32* | −0.62** | −0.43*** | −0.45 | −1.23** | −0.84** |
| | (0.19) | (0.23) | (0.16) | (0.39) | (0.46) | (0.31) |
| Constant | 12.80*** | 23.97*** | 16.96*** | 15.02*** | 31.11*** | 21.68*** |
| | (1.18) | (1.84) | (0.92) | (2.04) | (3.11) | (1.54) |
| $R^2$ | 0.28 | 0.27 | 0.29 | 0.12 | 0.14 | 0.13 |
| No. observations | 2,543 | 1,745 | 4,288 | 1,211 | 1,002 | 2,214 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. (Papers with a minority—but positive—number of female authors are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.18: Table H.1, majority female-authored

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 1.81*** | 0.26 | 0.36* | 0.23* | 0.13* |
| | (0.69) | (0.16) | (0.19) | (0.13) | (0.07) |
| Female ratio for men ($\beta_1 + \beta_2$) | 0.52 | 0.11 | 0.10 | 0.03 | 0.02 |
| | (0.70) | (0.14) | (0.17) | (0.12) | (0.07) |
| Female ratio×male ($\beta_2$) | −1.29 | −0.15 | −0.26 | −0.21 | −0.11 |
| | (0.96) | (0.21) | (0.25) | (0.17) | (0.10) |
| Lagged score ($\beta_0$) | 0.04* | 0.03 | 0.03 | 0.03 | 0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.34 | 0.71 | 0.68 | 0.69 | 0.49 |
| *z-test for no serial correlation* | | | | | |
| Order 1 | −18.53 | −13.32 | −14.90 | −17.51 | −18.77 |
| Order 2 | 0.74 | −0.52 | 0.22 | 0.51 | 0.25 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Columns display estimates identical to those in Table H.1, except that female ratio has been replaced with a dummy variable equal to 1 if a weak majority (50% or more) of authors are female. Otherwise, it is 0. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *J.4 At least one female author*

TABLE J.19: Table 2, at least one female author

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.50 | 0.50 | 0.49 | 0.51 | 0.64* | 0.39 | 0.36 | 0.35 |
| | (0.37) | (0.38) | (0.39) | (0.38) | (0.36) | (0.38) | (0.37) | (0.44) |
| Flesch-Kincaid | 0.11 | 0.11 | 0.11 | 0.12 | 0.13* | 0.13 | 0.14* | 0.10 |
| | (0.07) | (0.08) | (0.08) | (0.07) | (0.07) | (0.08) | (0.08) | (0.08) |
| Gunning Fog | 0.20** | 0.20** | 0.20** | 0.20** | 0.21** | 0.21** | 0.19** | 0.14 |
| | (0.09) | (0.09) | (0.09) | (0.09) | (0.08) | (0.10) | (0.09) | (0.10) |
| SMOG | 0.12* | 0.12* | 0.13* | 0.13* | 0.14** | 0.13* | 0.11* | 0.09 |
| | (0.07) | (0.07) | (0.07) | (0.07) | (0.06) | (0.07) | (0.06) | (0.07) |
| Dale-Chall | 0.07** | 0.07** | 0.07** | 0.08** | 0.08** | 0.08* | 0.07* | 0.07 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.04) | (0.04) | (0.04) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 9,117 articles in (1)–(5); 5,211 articles in (6) and (7); 5,774 articles—including 563 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.20: Table 4, at least one female author

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 0.63* | 1.36** | 1.76*** | 0.40 | 0.40 |
| | (0.34) | (0.58) | (0.67) | (0.27) | (0.28) |
| Flesch-Kincaid | 0.26** | 0.27** | 0.46*** | 0.20** | 0.20** |
| | (0.11) | (0.13) | (0.15) | (0.09) | (0.09) |
| Gunning Fog | 0.24** | 0.33** | 0.50*** | 0.17* | 0.17* |
| | (0.12) | (0.14) | (0.17) | (0.09) | (0.10) |
| SMOG | 0.16** | 0.21** | 0.33*** | 0.11* | 0.11* |
| | (0.08) | (0.09) | (0.11) | (0.06) | (0.06) |
| Dale-Chall | 0.10*** | 0.18*** | 0.25*** | 0.07*** | 0.07*** |
| | (0.03) | (0.05) | (0.06) | (0.02) | (0.02) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,709 NBER working papers; 1,707 published articles (377 female-authored). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.21: Table 6, at least one female author

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch Reading Ease | 0.46 | 0.33 | 3.09*** | 1.87* | 0.95 | 0.87 |
| | (0.41) | (0.56) | (0.83) | (1.05) | (1.35) | (0.56) |
| Flesch-Kincaid | 0.09 | −0.09 | 0.57*** | 0.54** | 0.18 | 0.09 |
| | (0.09) | (0.15) | (0.15) | (0.22) | (0.29) | (0.12) |
| Gunning Fog | 0.18* | 0.01 | 0.74*** | 0.69*** | 0.23 | 0.19 |
| | (0.11) | (0.18) | (0.18) | (0.26) | (0.34) | (0.14) |
| SMOG | 0.12 | 0.02 | 0.48*** | 0.47** | 0.19 | 0.15 |
| | (0.08) | (0.12) | (0.13) | (0.19) | (0.25) | (0.10) |
| Dale-Chall | 0.07** | 0.03 | 0.22*** | 0.16* | 0.12 | 0.09* |
| | (0.03) | (0.06) | (0.08) | (0.09) | (0.14) | (0.05) |
| No. observations | 6,875 | 2,826 | 1,675 | 1,906 | 2,773 | 12,006 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.22: Table 8 (dependent variable: revision duration), at least one female author

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1+ female | 2.721*** | 3.028*** | 3.019*** | 2.761*** | 3.032*** | 4.300*** | 4.298*** |
| | (0.932) | (0.999) | (1.001) | (1.000) | (0.999) | (1.173) | (1.146) |
| Max. $t_5$ | −0.138*** | −0.141*** | −0.140*** | −0.139*** | −0.139*** | −0.138*** | −0.145*** |
| | (0.037) | (0.037) | (0.037) | (0.038) | (0.037) | (0.043) | (0.044) |
| No. pages | 0.197*** | 0.197*** | 0.196*** | 0.197*** | 0.196*** | 0.239*** | 0.226*** |
| | (0.028) | (0.028) | (0.028) | (0.028) | (0.028) | (0.040) | (0.043) |
| $N_j$ | 1.016*** | 0.981*** | 0.969*** | 1.009*** | 0.977*** | 1.159*** | 1.033** |
| | (0.304) | (0.298) | (0.300) | (0.294) | (0.299) | (0.398) | (0.420) |
| Order | 0.207*** | 0.206*** | 0.204*** | 0.207*** | 0.204*** | 0.470*** | 0.468*** |
| | (0.063) | (0.063) | (0.063) | (0.064) | (0.063) | (0.144) | (0.142) |
| No. citations (asinh) | −0.395* | −0.411** | −0.400** | −0.394* | −0.411** | −0.633 | −0.640 |
| | (0.197) | (0.196) | (0.197) | (0.197) | (0.196) | (0.404) | (0.389) |
| Flesch Reading Ease | −0.017 | −0.016 | −0.016 | −0.017 | −0.016 | −0.035 | −0.037 |
| | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) | (0.027) | (0.028) |
| Mother | | | −4.147 | | −8.478*** | −17.264*** | −17.212*** |
| | | | (2.827) | | (3.123) | (5.288) | (5.495) |
| Birth | | | | −0.966 | 7.214 | 15.925** | 15.581** |
| | | | | (4.062) | (4.920) | (6.370) | (6.599) |
| Constant | 14.100*** | 14.199*** | 14.212*** | 14.114*** | 14.225*** | 16.689*** | 17.517*** |
| | (1.190) | (1.211) | (1.208) | (1.199) | (1.205) | (2.428) | (2.294) |
| $R^2$ | 0.287 | 0.288 | 0.287 | 0.287 | 0.288 | 0.122 | 0.140 |
| No. observations | 2,622 | 2,607 | 2,622 | 2,622 | 2,622 | 1,278 | 1,278 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.23: Table 9 (dependent variable: revision duration), at least one female author

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| 1+ female | 2.53** | 0.77 | 1.64*** | 3.79*** | 1.55* | 2.65*** |
| | (0.95) | (0.72) | (0.59) | (1.13) | (0.85) | (0.69) |
| Max. $t_5$ | −0.13*** | −0.12* | −0.13*** | −0.14*** | −0.07 | −0.11** |
| | (0.04) | (0.06) | (0.04) | (0.04) | (0.07) | (0.04) |
| No. pages | 0.20*** | 0.14** | 0.18*** | 0.22*** | 0.06 | 0.17*** |
| | (0.03) | (0.06) | (0.03) | (0.04) | (0.07) | (0.03) |
| $N_j$ | 1.02*** | −0.16 | 0.61** | 1.03** | 0.02 | 0.77* |
| | (0.31) | (0.49) | (0.28) | (0.44) | (0.66) | (0.38) |
| Order | 0.19*** | −0.07 | 0.08 | 0.46*** | 0.04 | 0.22* |
| | (0.06) | (0.08) | (0.05) | (0.13) | (0.15) | (0.12) |
| No. citations (asinh) | −0.30 | −0.64*** | −0.42*** | −0.42 | −1.20** | −0.83*** |
| | (0.19) | (0.23) | (0.15) | (0.38) | (0.44) | (0.30) |
| Constant | 12.97*** | 24.01*** | 17.12*** | 15.33*** | 31.07*** | 22.01*** |
| | (1.16) | (1.81) | (0.91) | (2.02) | (3.01) | (1.45) |
| $R^2$ | 0.28 | 0.27 | 0.29 | 0.12 | 0.14 | 0.13 |
| No. observations | 2,622 | 1,812 | 4,434 | 1,278 | 1,068 | 2,347 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.24: Table H.1, at least one female author

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 1.15* | 0.17 | 0.25 | 0.18 | 0.12* |
| | (0.67) | (0.14) | (0.17) | (0.12) | (0.06) |
| Female ratio for men ($\beta_1 + \beta_2$) | −0.02 | 0.01 | 0.03 | 0.02 | 0.05 |
| | (0.54) | (0.12) | (0.14) | (0.10) | (0.05) |
| Female ratio×male ($\beta_2$) | −1.18 | −0.16 | −0.23 | −0.16 | −0.07 |
| | (0.81) | (0.17) | (0.20) | (0.15) | (0.07) |
| Lagged score ($\beta_0$) | 0.04* | 0.05** | 0.03 | 0.03 | 0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.36 | 0.72 | 0.68 | 0.71 | 0.48 |
| *z-test for no serial correlation* | | | | | |
| Order 1 | −18.54 | −13.74 | −14.97 | −17.57 | −18.76 |
| Order 2 | 0.69 | −0.18 | 0.26 | 0.50 | 0.26 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Columns display estimates identical to those in Table H.1, except that female ratio has been replaced with a dummy variable equal to 1 if at least one author on a paper is female. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *J.5 Exclusively female-authored*

TABLE J.25: Table 2, 100% female-authored

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.49 | 0.41 | 0.34 | 0.51 | 0.79 | 0.52 | 0.61 | 0.75 |
| | (0.57) | (0.56) | (0.57) | (0.60) | (0.60) | (0.66) | (0.71) | (0.90) |
| Flesch-Kincaid | 0.11 | 0.09 | 0.08 | 0.13 | 0.16 | 0.21 | 0.24 | 0.21 |
| | (0.13) | (0.13) | (0.13) | (0.14) | (0.14) | (0.15) | (0.15) | (0.18) |
| Gunning Fog | 0.22 | 0.21 | 0.21 | 0.27* | 0.31* | 0.41** | 0.41** | 0.35* |
| | (0.14) | (0.14) | (0.14) | (0.15) | (0.16) | (0.17) | (0.17) | (0.20) |
| SMOG | 0.15 | 0.14 | 0.14 | 0.17 | 0.21* | 0.25** | 0.25** | 0.22 |
| | (0.10) | (0.10) | (0.10) | (0.11) | (0.11) | (0.12) | (0.12) | (0.15) |
| Dale-Chall | 0.06 | 0.06 | 0.05 | 0.06 | 0.08 | 0.13** | 0.13* | 0.15** |
| | (0.05) | (0.05) | (0.05) | (0.06) | (0.06) | (0.06) | (0.07) | (0.07) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 8,260 articles in (1)–(5); 4,455 articles in (6) and (7); 4,840 articles—including 385 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.26: Table 4, 100% female-authored

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 0.44 | 1.97 | 2.09 | 0.12 | 0.12 |
| | (0.95) | (1.38) | (1.70) | (0.97) | (0.99) |
| Flesch-Kincaid | 0.37 | 0.18 | 0.51 | 0.33 | 0.33 |
| | (0.28) | (0.24) | (0.38) | (0.30) | (0.31) |
| Gunning Fog | 0.41 | 0.33 | 0.67* | 0.34 | 0.34 |
| | (0.27) | (0.29) | (0.39) | (0.31) | (0.32) |
| SMOG | 0.19 | 0.33 | 0.45* | 0.12 | 0.12 |
| | (0.16) | (0.24) | (0.26) | (0.19) | (0.19) |
| Dale-Chall | 0.11 | 0.34** | 0.40*** | 0.06 | 0.06 |
| | (0.08) | (0.16) | (0.14) | (0.09) | (0.09) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 1,386 NBER working papers; 1,384 published articles (54 female-authored). Columns display estimates identical to those in Table 4, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.27: Table 6, 100% female-authored

| | $t_4 = 1$ | $t_4 = 2$ | $t_4 = 3$ | $t_4 = 4$–$5$ | $t_4 \geq 6$ | All |
|---|---|---|---|---|---|---|
| Flesch Reading Ease | −0.18 | 1.66 | 4.71*** | 1.97 | 1.77 | 3.08*** |
| | (0.80) | (1.09) | (1.21) | (2.55) | (3.28) | (1.19) |
| Flesch-Kincaid | −0.01 | 0.25 | 1.01*** | 0.07 | 0.45 | 0.47 |
| | (0.19) | (0.23) | (0.31) | (0.58) | (0.55) | (0.28) |
| Gunning Fog | 0.07 | 0.47* | 1.50*** | 0.43 | 0.87* | 0.72** |
| | (0.21) | (0.27) | (0.42) | (0.63) | (0.50) | (0.34) |
| SMOG | 0.02 | 0.33* | 1.01*** | 0.49 | 0.56* | 0.50** |
| | (0.15) | (0.20) | (0.29) | (0.51) | (0.30) | (0.23) |
| Dale-Chall | −0.02 | 0.19** | 0.38*** | 0.42 | 0.64*** | 0.25** |
| | (0.08) | (0.09) | (0.14) | (0.28) | (0.21) | (0.12) |
| No. observations | 5,880 | 2,449 | 1,453 | 1,642 | 2,384 | 8,079 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[3] | ✓[1] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*. Columns display estimates identical to those in Table 6, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.28: Table 8 (dependent variable: revision duration), 100% female-authored

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Exclusively female | 6.382** | 9.611** | 9.540** | 7.284** | 9.574** | 11.787** | 12.165** |
| | (2.641) | (3.619) | (3.621) | (3.198) | (3.620) | (4.999) | (4.986) |
| Max. $t_5$ | −0.174*** | −0.173*** | −0.172*** | −0.174*** | −0.170*** | −0.173*** | −0.169*** |
| | (0.045) | (0.045) | (0.045) | (0.045) | (0.045) | (0.056) | (0.058) |
| No. pages | 0.210*** | 0.208*** | 0.207*** | 0.209*** | 0.207*** | 0.254*** | 0.242*** |
| | (0.027) | (0.026) | (0.027) | (0.027) | (0.027) | (0.038) | (0.038) |
| $N_j$ | 1.267*** | 1.294*** | 1.275*** | 1.263*** | 1.285*** | 1.633*** | 1.441*** |
| | (0.319) | (0.317) | (0.321) | (0.318) | (0.319) | (0.436) | (0.427) |
| Order | 0.177*** | 0.175*** | 0.173*** | 0.176*** | 0.173*** | 0.422** | 0.409** |
| | (0.065) | (0.064) | (0.064) | (0.065) | (0.064) | (0.157) | (0.158) |
| No. citations (asinh) | −0.430** | −0.454** | −0.440** | −0.426** | −0.452** | −0.767* | −0.688 |
| | (0.194) | (0.194) | (0.194) | (0.195) | (0.193) | (0.423) | (0.419) |
| Flesch Reading Ease | −0.016 | −0.015 | −0.015 | −0.016 | −0.014 | −0.034 | −0.035 |
| | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.033) | (0.032) |
| Mother | | | −10.086** | | −14.522*** | −23.032*** | −23.988*** |
| | | | (4.444) | | (4.518) | (6.575) | (7.430) |
| Birth | | | | −4.903 | 7.360 | 15.319*** | 16.065*** |
| | | | | (5.082) | (4.586) | (5.428) | (5.645) |
| Constant | 13.739*** | 13.777*** | 13.791*** | 13.747*** | 13.803*** | 16.570*** | 17.017*** |
| | (1.316) | (1.336) | (1.330) | (1.322) | (1.327) | (3.032) | (2.782) |
| $R^2$ | 0.291 | 0.295 | 0.293 | 0.291 | 0.294 | 0.133 | 0.150 |
| No. observations | 2,443 | 2,428 | 2,443 | 2,443 | 2,443 | 1,140 | 1,140 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.29: Table 9 (dependent variable: revision duration), 100% female-authored

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Exclusively female | 6.42** | 2.46 | 4.22*** | 9.04** | 4.62* | 6.27*** |
| | (2.62) | (1.58) | (1.36) | (3.75) | (2.40) | (1.93) |
| Max. $t_5$ | −0.17*** | −0.12* | −0.15*** | −0.16*** | −0.07 | −0.13*** |
| | (0.04) | (0.07) | (0.04) | (0.05) | (0.08) | (0.05) |
| No. pages | 0.21*** | 0.13* | 0.19*** | 0.24*** | 0.03 | 0.19*** |
| | (0.03) | (0.08) | (0.03) | (0.04) | (0.09) | (0.03) |
| $N_j$ | 1.30*** | 0.07 | 0.87*** | 1.38*** | 0.65 | 1.19*** |
| | (0.33) | (0.50) | (0.29) | (0.44) | (0.69) | (0.37) |
| Order | 0.17** | −0.03 | 0.09 | 0.41** | 0.14 | 0.25** |
| | (0.06) | (0.08) | (0.05) | (0.15) | (0.16) | (0.12) |
| No. citations (asinh) | −0.34* | −0.45* | −0.38** | −0.47 | −0.95* | −0.76** |
| | (0.18) | (0.23) | (0.15) | (0.39) | (0.47) | (0.33) |
| Constant | 12.58*** | 22.38*** | 16.20*** | 14.86*** | 28.93*** | 20.58*** |
| | (1.28) | (1.77) | (0.99) | (2.50) | (3.10) | (1.80) |
| $R^2$ | 0.29 | 0.28 | 0.30 | 0.13 | 0.16 | 0.15 |
| No. observations | 2,443 | 1,636 | 4,079 | 1,140 | 914 | 2,055 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Papers written by authors of both genders are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.30: Table H.1, 100% female-authored

| | Flesch Reading Ease | Flesch-Kincaid | Gunning Fog | SMOG | Dale-Chall |
|---|---|---|---|---|---|
| Female ratio for women ($\beta_1$) | 1.71 | 0.11 | 0.41 | 0.36* | 0.31*** |
| | (1.04) | (0.23) | (0.28) | (0.20) | (0.10) |
| Lagged score ($\beta_0$) | 0.04** | 0.05** | 0.03 | 0.03 | 0.03 |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| Hansen test ($p$-value) | 0.30 | 0.68 | 0.56 | 0.67 | 0.42 |
| *$z$-test for no serial correlation* | | | | | |
| Order 1 | −18.52 | −13.74 | −14.96 | −17.56 | −18.64 |
| Order 2 | 0.72 | −0.18 | 0.27 | 0.51 | 0.12 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | ✓[2] | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes.* Sample 9,180 observations (2,826 authors, 121 female). Estimates identical to those in Table H.1, except that female ratio has been replaced with a dummy variable equal to 1 if all authors on a paper are female. (Co-authored mixed-sex papers are included and classified as male (see Footnote 46).) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

### *J.6 Inexperienced senior female author*

TABLE J.31: Table 2, inexperienced senior female author

| | 1950–2015 | | | | | 1990–2015 | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flesch Reading Ease | 0.67 | 0.66 | 0.53 | 0.63 | 0.73 | 0.35 | 0.40 | 0.20 |
| | (0.54) | (0.51) | (0.52) | (0.55) | (0.58) | (0.65) | (0.63) | (0.86) |
| Flesch-Kincaid | 0.11 | 0.10 | 0.08 | 0.12 | 0.13 | 0.14 | 0.17 | 0.13 |
| | (0.15) | (0.14) | (0.15) | (0.15) | (0.15) | (0.17) | (0.16) | (0.20) |
| Gunning Fog | 0.19 | 0.19 | 0.17 | 0.22 | 0.23 | 0.27 | 0.26 | 0.22 |
| | (0.16) | (0.15) | (0.15) | (0.15) | (0.16) | (0.18) | (0.17) | (0.20) |
| SMOG | 0.11 | 0.11 | 0.10 | 0.13 | 0.15 | 0.15 | 0.14 | 0.12 |
| | (0.11) | (0.10) | (0.10) | (0.10) | (0.11) | (0.12) | (0.11) | (0.14) |
| Dale-Chall | 0.04 | 0.05 | 0.04 | 0.04 | 0.05 | 0.06 | 0.06 | 0.06 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.07) |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Journal effects | ✓ | ✓ | | | | | | |
| Year effects | | ✓ | | | | | | |
| Journal×Year effects | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $N_j$ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quality controls | | | | | ✓[1] | ✓[1] | ✓[1] | ✓[1] |
| Native speaker | | | | | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ | |
| *JEL* (tertiary) effects | | | | | | | | ✓ |

*Notes.* 4,984 articles in (1)–(5); 2,163 articles in (6) and (7); 2,411 articles—including 248 from *AER Papers & Proceedings* (see Footnote 12)—in (8). Estimates are identical to those in Table 2, except that only papers by junior authors (defined as having two or fewer previous top-five articles) are included in the sample and female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.32: Table 4, inexperienced senior female author

| | OLS | FGLS | | | OLS |
|---|---|---|---|---|---|
| | Published article | Working paper | Published article | Difference | Change in score |
| Flesch Reading Ease | 1.97* | 0.99 | 2.74 | 1.75 | 1.75 |
| | (1.12) | (1.67) | (1.98) | (1.12) | (1.19) |
| Flesch-Kincaid | 0.76** | 0.11 | 0.84** | 0.72** | 0.72** |
| | (0.32) | (0.34) | (0.42) | (0.32) | (0.34) |
| Gunning Fog | 0.77** | 0.12 | 0.85** | 0.74** | 0.74** |
| | (0.32) | (0.38) | (0.43) | (0.35) | (0.37) |
| SMOG | 0.46** | 0.12 | 0.55* | 0.43* | 0.43* |
| | (0.20) | (0.27) | (0.29) | (0.22) | (0.23) |
| Dale-Chall | 0.18 | 0.22 | 0.35* | 0.13 | 0.13 |
| | (0.12) | (0.16) | (0.18) | (0.11) | (0.11) |
| Editor effects | ✓ | ✓ | ✓ | | ✓ |
| Journal×Year effects | ✓ | ✓ | ✓ | | ✓ |
| $N_j$ | ✓ | ✓ | ✓ | | ✓ |
| Quality controls | ✓[2] | ✓[2] | ✓[2] | | ✓[2] |
| Native speaker | ✓ | ✓ | ✓ | | ✓ |

*Notes.* Sample 473 NBER working papers; 472 published articles (54 female-authored). Columns display estimates identical to those in Table 4, except that only papers by junior authors (defined as having two or fewer previous top-five articles) are included in the sample and female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.33: Table 8 (dependent variable: revision duration), inexperienced senior female author

| | 1970–2015 | | | | | 1990–2015 | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Senior female | 5.261* | 7.476** | 7.391** | 6.206* | 7.402** | 9.200 | 9.003 |
| | (2.899) | (3.621) | (3.604) | (3.293) | (3.608) | (5.485) | (5.672) |
| Max. $t_5$ | 0.702* | 0.672* | 0.681* | 0.679* | 0.688* | 1.031 | 0.897 |
| | (0.385) | (0.385) | (0.383) | (0.387) | (0.384) | (0.720) | (0.774) |
| No. pages | 0.174*** | 0.177*** | 0.171*** | 0.173*** | 0.171*** | 0.175*** | 0.181*** |
| | (0.038) | (0.037) | (0.037) | (0.037) | (0.037) | (0.059) | (0.057) |
| $N_j$ | −0.039 | −0.024 | −0.045 | −0.040 | −0.046 | −0.268 | −0.258 |
| | (0.484) | (0.487) | (0.485) | (0.483) | (0.485) | (0.999) | (1.112) |
| Order | 0.040 | 0.038 | 0.038 | 0.038 | 0.038 | 0.145 | 0.041 |
| | (0.081) | (0.081) | (0.081) | (0.081) | (0.081) | (0.234) | (0.267) |
| No. citations (asinh) | −0.327 | −0.359 | −0.339 | −0.318 | −0.348 | 0.010 | 0.072 |
| | (0.218) | (0.218) | (0.218) | (0.220) | (0.219) | (0.707) | (0.721) |
| Flesch Reading Ease | 0.018 | 0.019 | 0.019 | 0.018 | 0.019 | −0.012 | −0.006 |
| | (0.021) | (0.021) | (0.021) | (0.021) | (0.021) | (0.055) | (0.055) |
| Mother | | | −9.097** | | −11.534** | −18.565* | −19.418* |
| | | | (4.402) | | (4.476) | (9.355) | (9.727) |
| Birth | | | | −6.584 | 3.899 | 6.811 | 8.006 |
| | | | | (5.008) | (4.224) | (8.265) | (9.233) |
| Constant | 12.483*** | 12.519*** | 12.607*** | 12.528*** | 12.614*** | 17.258*** | 17.551*** |
| | (1.486) | (1.498) | (1.481) | (1.492) | (1.478) | (3.567) | (3.843) |
| $R^2$ | 0.327 | 0.332 | 0.330 | 0.328 | 0.330 | 0.138 | 0.180 |
| No. observations | 1,340 | 1,329 | 1,340 | 1,340 | 1,340 | 468 | 464 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Institution effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *JEL* (primary) effects | | | | | | | ✓ |

*Notes.* Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 8, except that only papers by junior authors (defined as having two or fewer previous top-five articles) are included in the sample and female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

TABLE J.34: Table 9 (dependent variable: revision duration), inexperienced senior female author

| | 1970–2015 | | | 1990–2015 | | |
|---|---|---|---|---|---|---|
| | *Econometrica* | *REStud* | *Econometrica* + *REStud* | *Econometrica* | *REStud* | *Econometrica* + *REStud* |
| Senior female | 4.92* | 2.70 | 3.72** | 6.24 | 4.10* | 5.07** |
| | (2.86) | (1.76) | (1.48) | (4.65) | (2.32) | (2.09) |
| Max. $t_5$ | 0.66* | 0.75 | 0.69*** | 0.65 | 0.59 | 0.72 |
| | (0.37) | (0.48) | (0.25) | (0.79) | (0.72) | (0.52) |
| No. pages | 0.17*** | 0.18* | 0.18*** | 0.18*** | 0.05 | 0.15*** |
| | (0.04) | (0.10) | (0.04) | (0.05) | (0.12) | (0.05) |
| $N_j$ | −0.13 | −1.57*** | −0.67* | −0.52 | −1.62* | −0.93 |
| | (0.47) | (0.54) | (0.40) | (1.04) | (0.82) | (0.64) |
| Order | 0.03 | 0.02 | 0.04 | 0.14 | 0.18 | 0.16 |
| | (0.08) | (0.07) | (0.05) | (0.26) | (0.16) | (0.15) |
| No. citations (asinh) | −0.32 | −0.52* | −0.40** | 0.02 | −1.46** | −0.88* |
| | (0.22) | (0.28) | (0.17) | (0.73) | (0.62) | (0.50) |
| Constant | 13.42*** | 21.38*** | 16.63*** | 17.76*** | 32.54*** | 24.71*** |
| | (1.37) | (2.13) | (1.18) | (3.58) | (3.99) | (2.49) |
| $R^2$ | 0.32 | 0.34 | 0.36 | 0.15 | 0.19 | 0.15 |
| No. observations | 1,340 | 1,006 | 2,346 | 464 | 499 | 965 |
| Editor effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Accepted year effects | ✓ | ✓ | | ✓ | ✓ | |
| Journal×Accepted year effects | | | ✓ | | | ✓ |
| *JEL* (primary) effects | | | | ✓ | ✓ | ✓ |

*Notes*. Dependent variable is the number of months spent in peer review (see Section 3.4). Columns display estimates identical to those in Table 9, except that only papers by junior authors (defined as having two or fewer previous top-five articles) are included in the sample and female ratio has been replaced with a dummy variable equal to 1 if a female author had at least as many top-five papers as her co-authors at the time the paper was published. (Mixed-gendered papers with a senior male co-author are excluded.) ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

## K Discussion of potential alternative explanations

A gender readability gap exists. It's still there after including editor, journal and year effects—meaning it's hard blame specific policies or attitudes in the fifties, long since overcome. The gap is unaffected by field controls—*i.e.*, it unlikely results from women researching topics that are easier to explain. Nor does it appear to be caused by factors correlated with gender but actually linked to authors' (or co-authors') competence as economists and fluency in English—if so, institution, native speaker and citation controls would reduce it. They do not.[48]

The gap grows between first draft and final publication and over the course of women's careers, precluding inborn advantage and one-off improvements in response to external circumstances unrelated to peer review. This likewise rules out gender differences in (i) biology/behaviour—*e.g.*, sensitivity to referee criticism[49]—or (ii) knowledge about referee expectations. If diligently addressing every referee concern has no apparent upside—acceptance rates are unaffected—and a very clear downside—constant redrafting takes time—even the most oversensitive, ill-informed woman would *eventually* re-examine initial beliefs and start acting like a man.[50] Yet this is not what we observe. The largest investments in writing well are made by female economists with greatest exposure to peer review—*i.e.*, those with the best opportunity to update their priors.

Women's papers are more likely assigned female referees (Abrevaya and Hamermesh, 2012; Gilbert et al., 1994).[51] If women are more demanding critics, clearer writing could reflect their tougher reviews.[52] Women concentrate in particular fields, so it's natural female referees more often review female-authored papers. Nevertheless, for the readability gap to exist only because of specialisation, controlling for *JEL* classification should explain it.[53] It does not: including 20 primary or 731 tertiary *JEL* category dummies has little effect. So if referee assignment is causing the gap, it's only because journals disproportionately refer female-authored papers to the toughest critics.[54] Meaning it isn't referees who are biased—it's editors.[55]

Section 3.2 directly links an increase in the gender readability gap to peer review; Section 3.3 establishes that factors outside women's control—assumed, at this point, entirely peer-review-related—drive it. Yet oversensitivity and/or poor information could create the former gap while *another* gender bias unconnected to peer review generates the latter. One in particular comes to mind: the feedback

[48] I also conducted a primitive surname analysis (see Hengel, 2016, pp. 35–36). It suggests that the female authors in my data are no more or less likely to be native English speakers.

[49] While women do appear more *internally* responsive to feedback—criticism has a bigger impact on their self-esteem—available evidence suggests they aren't any more *externally* responsive to it, *i.e.*, women and men are equally likely to change behaviour and alter performance after receiving feedback (Johnson and Helgeson, 2002; Roberts and Nolen-Hoeksema, 1989).

[50] This statement is especially relevant if the opportunity cost to women for "wasting" time on needless tasks is higher—*e.g.*, because of family responsibilities.

[51] Note that women are only a fraction of all referees—8 percent in 1986 (Blank, 1991), 10 percent in 1994 (Hamermesh, 1994) and 14 percent in 2013 (Torgler and Piatti, 2013). Abrevaya and Hamermesh (2012) report female-authored papers were only slightly more likely to be assigned a female referee between 1986–1994; matching increases between 2000–2008.

[52] It's not clear whether women's reports are more critical. A study specific to post-graduate biologists suggests yes (Borsuk et al., 2009); another analysing past reviews in an economics field journal does not (Abrevaya and Hamermesh, 2012).

[53] Specifically, men and women publishing in the same field face the same pool of referees. Controlling for that pool would account for gender differences in readability.

[54] Relatedly, perhaps female-authored research is more provocative and therefore warrants more scrutiny. Yet if this explained the gap, controlling for *JEL* classification should reduce (or eliminate) it—unless women's work is systematically more provocative even among researchers in very narrow fields. There is some evidence for this hypothesis—provocative work is (presumably) highly cited work and recent female-authored papers published in top economics journals are cited more (**Hengel2018a**). Yet more provocative, cited research would probably be published at higher rates—and there is no evidence women's paper's are more frequently accepted (Ceci et al., 2014). In any case, women respond to incentives just like men; if we could get boring papers published, we'd write them.

[55] This is a form of biased referee assignment identified in Theorem 1. It would also apply if the readability gap reflects referees' apathy for women's work. Readability is particularly relevant when interest in—and knowledge about—the topic is low (Fass and Schumacher, 1978; Klare, 1976). Thus, a gap could emerge if editors fail to assign interested and knowledgable referees to female-authored papers.

women receive in conferences and seminars. Perhaps experienced female economists tighten prose (before or after submission) in response to audience member remarks.

Anecdotal evidence suggests female speakers are given a harder time, although I could find no scientific analysis to support (or refute) this claim. Nevertheless, sensible, experienced economists should ignore random suggestions that won't actually improve a manuscript's probability of acceptance. Do well-published female economists really lack this sensibility? In any case, most conference and seminar participants are also current (or future) journal referees. Neutral peer review feedback is inconsistent with non-neutral conference/seminar feedback when originating from the same group—especially since gender neutrality is emphasised in both environments.

Perhaps women focus on writing at the expense of some other aspect of a paper due to a comparative advantage? Not likely. Women's chosen publication strategy results in similar (or lower) acceptance rates and longer review times compared to the one employed by men. But if men and women are equally capable researchers then writing well cannot be a comparative advantage and at the same time be strictly dominated by another strategy.[56]

In the universe of straightforward alternatives, this leaves us with one: female economists are less capable researchers. As mentioned earlier, factors correlated with gender but actually related to competency should decline when appropriate proxies are included. The sample itself is one such proxy—these are, after all, only articles published in the top four economics journals. Adding other controls—author seniority, institution, total article count, citations and published order in an issue—has no effect.[57] The gap is widest for the most productive economists and even exists among articles originally released as NBER working papers—both presumably very clear signals of merit. Indeed, contemporary female-authored papers published in a top-four economics journal are, in fact, cited more than male-authored papers (Hengel and Moon, 2019).

Yet I cannot rule out the possibility that women's work is systematically worse than men's in a way that is somehow not full captured by citations, proxies for author prominence and seniority or author-specific fixed effects—or that the female and male authors in Section 3.3.2 are not really equivalent. (To decide for yourself, see Appendix F.4.) And if this is true, editors and referees *should* select and peruse our papers more carefully—a byproduct of which could be better written papers after-the-fact or more attractive prose compensating for structural weaknesses before it.

"Quality" is subjective; measurement, not easy. Nevertheless, attempts using citation counts and journal acceptance rates do not indicate that men's research is any better: as discussed in Section 3.3.1, gender has very little impact on the latter;[58] a review of past studies on male vs. female citations find four in which women's papers received fewer, six where they were cited more and eight with no significant difference (Ceci et al., 2014). Recent research specific to economics suggests *female-authored papers get cited more* (Card et al., 2019; Grossbard et al., 2018; Hengel and Moon, 2019).

More complicated, multi-factor explanations could resolve inconsistencies present when each is analysed in isolation. Perhaps female economists are perfectionists, and it gets stronger with age?[59] Maybe women actually enjoy being poorly informed, overconfident and sensitive to criticism—or (more likely) I could have otherwise misspecified the author's objective function in Section 3.3.1. Meanwhile, a preference for writing well coupled with unaccounted for exogenously determined co-author characteristics could combine to cause women's more readable papers *and* their increasing readabil-

[56] Assuming men and women are equally capable researchers, women would only emphasise a particular aspect of a paper at the expense of others if doing so achieved a similar outcome/effort trade-off as the one employed by men. The outcome/effort combination women *currently* experience, however, is strictly worse than men's.

[57] Published order in an issue was introduced as a set of indicator variables in an earlier version of this paper (Hengel, 2016, pp. 42 and 44).

[58] Journals may have a policy of publishing female-authored research over equal (or even better) male work. If so, acceptance rates are not an unbiased indicator of quality.

[59] While women score higher on maintaining order (Feingold, 1994)—a trait including organisation and perfectionism—significant differences are not universally present in all cultures (Costa et al., 2001); differences that are present decline—or even reverse—as people age (Weisberg et al., 2011).

ity[60]—although restricting the analysis to solo-authored papers, those co-authored by members of the same sex or with a senior female co-author results in similar figures and identical conclusions (see Appendix J.5, Table J.27 and the robustness discussion in Section 3.3.2).[61] Alternatively, measurement error and/or co-variate controls could have interacted with gender in ways I did not anticipate.[62] And of course, the statistically significant relationships this paper documents may simply be unfortunate (particularly for me!) flukes.[63]

Still, no explanation matches the simplicity and believability of biased referees and/or editors. Coherence and economy do not establish fact, but they are useful guides. This single explanation neatly accounts for all observed patterns. If reviewers apply higher standards to female-authored papers, they will be rejected more often and/or subject to tougher review. Added scrutiny should improve exposition but prolong publication. The rewards from clearer writing are presumably internalised, explaining gradual increases in women's readability.

Moreover, several studies document a gender difference in critical feedback of similar form—employee performance reviews and student evaluations. Ongoing research suggests female workers are held to higher standards in job assessments. They are acknowledged less for creativity and technical expertise, their contributions are infrequently connected to business outcomes; guidance or praise supervisors do offer is vague (Correll and Simard, 2016).[64]

Students display a similar bias. Data from Rate My Professors suggest female lecturers should be "helpful", "clear", "organised" and "friendly". Men, instead, are praised (and criticised) for being "smart", "humble" or "cool" (Schmidt, 2015).[65] A study of teaching evaluations similarly finds students value preparation, organisation and clarity in female instructors; their male counterparts are considered more knowledgable, praised for their "animation" and "leadership" and given more credit for contributing to students' intellectual development (Boring, 2017).

[60] This might occur if senior women are excluded from male networks as $t$ increases; consequently, they are more likely to co-author with other women than junior female economists. As I show in an earlier version of this paper, however, the reverse is true: as $t$ increases, women are more likely to co-author with men, while men are more likely to co-author with women (Hengel, 2016, Table 12, p. 25).

[61] Relatedly, women may have preferred to have written their $t_4 = 1$ publication more clearly, but senior male co-authors held them back; at $t_4 = 3$, they enjoy more freedom to achieve their desired (higher) readability by writing on their own or with other women. This runs counter to the observation in Footnote 60, however. Moreover, women are more likely to co-author with more senior men at $t_4 = 3$ than they were at $t_4 = 1$.

[62] Appendix B.3 outlines principle sources of measurement error as well as steps I have taken to minimise their impact. Meanwhile, coefficient magnitude and standard errors remain relatively stable when gradually introducing controls (Table 2), reducing the likelihood of "collider" bias (see Footnote 14).

[63] This is a form of "file drawer bias"—other studies showing no effect weren't published. Nevertheless, at least one recent paper found similar results: the readability of disclosure documents in audit reports is positively correlated with the proportion of women and underrepresented minorities on an audit committee (Velte, 2018).

[64] A similar phenomenon exists in online fora. The *Guardian* commissioned researchers to study 70 million comments on its website. It found female and black writers attract disproportionately abusive threads (Gardiner et al., 2016).

[65] These conclusions are based on my own observational account of the data.

## References


Abrevaya, J. and D. S. Hamermesh (2012). "Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)?" *Review of Economics and Statistics* 94 (1), pp. 202–207 (cit. on p. 84).

Alkhurayyif, Y. and G. R. S. Weir (2017). "Readability as a Basis for Information Security Policy Assessment". In: *Seventh International Conference on Emerging Security Technologies*. Canterbury, pp. 114–121 (cit. on p. 8).

Arellano, M. and S. Bond (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". *Review of Economic Studies* 58 (2), pp. 277–297 (cit. on p. 44).

Arellano, M. and O. Bover (1995). "Another Look at the Instrumental Variable Estimation of Error-components Models". *Journal of Econometrics* 68 (1), pp. 29–51 (cit. on pp. 43, 44).

Armbruster, B. B. (1984). "The Problem of Inconsiderate Text". In: *Comprehension Instruction: Perspectives and Suggestions*. Ed. by G. G. Duffy, L. R. Roehler, and J. Mason. New York, New York: Longman, pp. 202–217 (cit. on p. 10).

Benoit, K., K. Munger, and A. Spirling (2017). "Measuring and Explaining Political Sophistication through Textual Complexity". Mimeo (cit. on p. 12).

Berninger, M. et al. (2017). "Confused but Convinced: Article Complexity and the Number of Citations". Mimeo (cit. on p. 7).

Biddle, G. C., G. Hilary, and R. S. Verdi (2009). "How Does Financial Reporting Quality Relate to Investment Efficiency?" *Journal of Accounting and Economics* 48 (2-3), pp. 112–131 (cit. on p. 8).

Blank, R. M. (1991). "The Effects of Double-blind versus Single-blind Reviewing: Experimental Evidence from the American Economic Review". *American Economic Review* 81 (5), pp. 1041–1067 (cit. on pp. 15, 84).

Blundell, R. and S. Bond (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". *Journal of Econometrics* 87 (1), pp. 115–143 (cit. on pp. 43, 44).

Boring, A. (2017). "Gender Biases in Student Evaluations of Teaching". *Journal of Public Economics* 145 (Supplement C), pp. 27–41 (cit. on p. 86).

Borsuk, R. M. et al. (2009). "To Name or Not to Name: The Effect of Changing Author Gender on Peer Review". *BioScience* 59 (11), pp. 985–989 (cit. on p. 84).

Card, D. et al. (2019). "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics* 135 (1), pp. 269–327 (cit. on p. 85).

Ceci, S. J. et al. (2014). "Women in Academic Science: A Changing Landscape". *Psychological Science in the Public Interest* 15 (3), pp. 75–141 (cit. on pp. 84, 85).

Chall, J. S., A. Freeman, and B. Levy (1983). "Minimum Competency Testing of Reading: An Analysis of Eight Tests Designed for Grade 11". In: *The Courts, Validity, and Minimum Competency Testing*. Ed. by G. F. Madaus. Boston, Massachussetts: Kluwer-Nijhoff. Chap. 10, pp. 197–208 (cit. on p. 8).

Chall, J. S., S. S. Conard, and S. H. Harris (1977). *An Analysis of Textbooks in Relation to Declining SAT Scores*. Tech. rep. Prepared fo the Advisory Panel on on the Scholastic Aptitude Test Score Decline. Princeton, New Jersey (cit. on p. 8).

Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: Brookline Books (cit. on pp. 7, 10, 11).

Coleman, E. B. (1964). "The Comprehensibility of Several Grammatical Transformations". *Journal of Applied Psychology* 48 (3), pp. 186–190 (cit. on p. 9).

— (1965). "Learning of Prose Written in Four Grammatical Transformations". *Journal of Applied Psychology* 49 (5), pp. 332–341 (cit. on p. 9).

Correll, S. and C. Simard (2016). "Vague Feedback Is Holding Women Back". *Harvard Business Review*. https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back. Accessed: 2016-10-04 (cit. on p. 86).

Costa, P. T., A. Terracciano, and R. R. McCrae (2001). "Gender Differences in Personality Traits Across Cultures: Robust and Surprising Findings". *Journal of Personality and Social Psychology* 81 (2), pp. 322–331 (cit. on p. 85).
Cox, C. (2007). *Closing Remarks to the Second Annual Corporate Governance Summit*. Delivered at USC Marshall School of Business, Los Angeles, California, 23 March (cit. on p. 8).
Dale, E. and J. S. Chall (1948). "A Formula for Predicting Readability". *Educational Research Bulletin* 27 (1), pp. 11–20 (cit. on p. 10).
De Franco, G. et al. (2015). "Analyst Report Readability". *Contemporary Accounting Research* 32 (1), pp. 76–104 (cit. on p. 8).
Dowling, M., H. Hammami, and O. Zreik (2018). "Easy to Read, Easy to Cite?" *Economics Letters* 173, pp. 100–103 (cit. on p. 7).
DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information (cit. on p. 7).
Enke, B. (2018). "Moral Values and Voting: Trump and Beyond". NBER Working Paper Series, No. 24268 (cit. on p. 8).
Fass, W. and G. M. Schumacher (1978). "Effects of Motivation, Subject Activity, and Readability on the Retention of Prose Materials". *Journal of Educational Psychology* 70 (5), pp. 803–807 (cit. on p. 84).
Feingold, A. (1994). "Gender Differences in Personality: A Meta-analysis". *Psychological Bulletin* 116 (3), pp. 429–456 (cit. on p. 85).
Flesch, R. (1949). *The Art of Readable Writing*. New York, New York: Harper and Brothers Publishers (cit. on p. 7).
Foster, D. R. and D. H. Rhoney (2002). "Readability of Printed Patient Information for Epileptic Patients". *Annals of Pharmacotherapy* 36 (12), pp. 1856–1861 (cit. on p. 8).
Gardiner, B. et al. (2016). "The Dark Side of Guardian Comments". *Guardian*. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments. Accessed: 2016-10-04 (cit. on p. 86).
Gilbert, J. R., E. S. Williams, and G. D. Lundberg (1994). "Is There Gender Bias in JAMA's Peer Review Process?" *Journal of the American Medical Association* 272 (2), pp. 139–142 (cit. on p. 84).
Grossbard, S., T. Yilmazer, and L. Zhang (2018). "The Gender Gap in Citations of Economics Articles: Lessons from Economics of the Household". Mimeo (cit. on p. 85).
Guerini, M., A. Pepe, and B. Lepri (2012). "Do Linguistic Style and Readability of Scientific Abstracts Affect their Virality?" In: *Proceedings of the Sixth Interntaional AAAI Conference of Weblogs and Social Media*. Dublin, pp. 475–478 (cit. on p. 7).
Hamermesh, D. S. (1994). "Facts and Myths about Refereeing". *Journal of Economic Perspectives* 8 (1), pp. 153–163 (cit. on p. 84).
Hartley, J., J. W. Pennebaker, and C. Fox (2003). "Using New Technology to Assess the Academic Writing Styles of Male and Female Pairs and Individuals". *Journal of Technical Writing and Communication* 33 (3), pp. 243–261 (cit. on p. 10).
Hengel, E. (2016). "Publishing while Female: Gender Differences in Peer Review Scrutiny". Mimeo (cit. on pp. 15, 18, 43, 84–86).
Hengel, E. and E. Moon (2019). "Gender and Quality at Top Economics Journals". Mimeo (cit. on p. 85).
Hussin, M. F. et al. (2012). "The Readability of Transmission Line Characteristics Lab Manual". In: *IEEE Control and System Graduate Research Colloquium*. Shah Alam, Selangor, pp. 398–401 (cit. on p. 8).
Jansen, D. J. (2011). "Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies". *Contemporary Economic Policy* 29 (4), pp. 494–509 (cit. on p. 8).
Johnson, M. and V. S. Helgeson (2002). "Sex Differences in Response to Evaluative Feedback: A Field Study". *Psychology of Women Quarterly* 26 (3), pp. 242–251 (cit. on p. 84).

Kemper, S. (1983). "Measuring the Inference Load of a Text". *Journal of Educational Psychology* 75 (3), pp. 391–401 (cit. on p. 10).
King, D. W., C. Tenopir, and M. Clarke (2006). "Measuring Total Reading of Journal Articles". *D-Lib Magazine* 12 (10), pp. 1082–9873 (cit. on p. 10).
Kintsch, W. and J. R. Miller (1984). "Readability: A View from Cognitive Psychology". In: *Understanding Reading Comprehension*. Ed. by J. Flood. Newark, Delaware: International Reading Association, pp. 220–232 (cit. on p. 10).
Klare, G. R. (1976). "Judging Readability". *Instructional Science* 5 (1), pp. 55–61 (cit. on p. 84).
Klare, G. R. and K. L. Smart (1973). "Analysis of the Readability Level of Selected USAFI Instructional Materials". *Journal of Educational Research* 67 (4), p. 176 (cit. on pp. 7, 8).
Kleven, H. J. (2018). *Language Trends in Public Economics*. https://www.henrikkleven.com/uploads/3/7/3/1/37310663/languagetrends_slides_kleven.pdf. Accessed: 2018-12-02 (cit. on p. 11).
Laband, D. N. and C. N. Taylor (1992). "The Impact of Bad Writing in Economics". *Economic Inquiry* 30 (4), pp. 673–688 (cit. on p. 7).
Law, D. S. and D. Zaring (2010). "Law versus Ideology: the Supreme Court and the Use of Legislative History". *William and Mary Law Review* 51 (5), pp. 1653–1747 (cit. on p. 8).
Lawrence, A. (2013). "Individual Investors and Financial Disclosure". *Journal of Accounting and Economics* 56 (1), pp. 130–147 (cit. on p. 8).
Lehavy, R., F. Li, and K. Merkley (2011). "The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts". *Accounting Review* 86 (3), pp. 1087–1115 (cit. on p. 8).
Lei, L. and S. Yan (2016). "Readability and Citations in Information Science: Evidence from Abstracts and Articles of Four Journals (2003–2012)". *Scientometrics* 108 (3), pp. 1155–1169 (cit. on p. 7).
Leuven, E. and B. Sianesi (2003). *PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing*. http://ideas.repec.org/c/boc/bocode/s432001.html. Accessed: 2018-03-20 (cit. on p. 28).
Li, F. (2008). "Annual Report Readability, Current Earnings, and Earnings Persistence". *Journal of Accounting and Economics* 45 (2-3), pp. 221–247 (cit. on p. 8).
Liang, F. M. (1983). "Word Hy-phen-a-tion by Com-put-er". PhD thesis. Stanford University (cit. on p. 11).
Loughran, T. and B. McDonald (2016). "Textual Analysis in Accounting and Finance: A Survey". *Journal of Accounting Research* 54 (4), pp. 1187–1230 (cit. on p. 8).
Marino Fages, D. (2020). "Write better , publish better". *Scientometrics* (forthcoming) (cit. on p. 7).
McCannon, B. C. (2019). "Readability and Research Impact". *Economics Letters* 180 (July), pp. 76–79 (cit. on p. 7).
Meade, C. D. and J. C. Byrd (1989). "Patient Literacy and the Readability of Smoking Education Literature". *American Journal of Public Health* 79 (2), pp. 204–206 (cit. on p. 8).
Meyer, B. J. F. (1982). "Reading Research and the Composition Teacher: The Importance of Plans". *College Composition and Communication* 33 (1), pp. 37–49 (cit. on p. 10).
Miller, B. P. (2010). "The Effects of Reporting Complexity on Small and Large Investor Trading". *Accounting Review* 85 (6), pp. 2107–2143 (cit. on p. 8).
Plavén-Sigray, P. et al. (2017). "The Readability of Scientific Texts is Decreasing over Time". *eLife* 6 (e27725), pp. 1–14 (cit. on p. 10).
Püttmann, L. (2017). *VoxEU Gobbledygook*. http://lukaspuettmann.com/2017/12/09/voxeu-gobbledygook/. Accessed: 2018-05-19 (cit. on p. 7).
Richardson, J. V. (1977). "Readability and Readership of Journals in Library Science". *Journal of Academic Librianship* 3 (1), pp. 20–22 (cit. on p. 7).
Roberts, T.-A. and S. Nolen-Hoeksema (1989). "Sex Differences in Reactions to Evaluative Feedback". *Sex Roles* 21 (11-12), pp. 725–747 (cit. on p. 84).
Sawyer, A. G., J. Laran, and J. Xu (2008). "The Readability of Marketing Journals: Are Award-Winning Articles Better Written?" *Journal of Marketing* 72 (1), pp. 108–117 (cit. on p. 7).

Schmidt, B. (2015). "Gender Bias Exists in Professor Evaluations". *New York Times*. http://www.nytimes.com/roomfordebate/2015/12/16/is-it-fair-to-rate-professors-online/gender-bias-exists-in-professor-evaluations. Accessed: 2016-10-04 (cit. on p. 86).
Sirico, L. J. (2007). "Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations". *Quinnipiac Law Review* 26 (1), pp. 147–172 (cit. on p. 11).
Spirling, A. (2016). "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915". *Journal of Politics* 78 (1), pp. 120–136 (cit. on p. 8).
Swanson, C. E. (1948). "Readability and Readership: A Controlled Experiment". *Journalism Bulletin* 25 (4), pp. 339–343 (cit. on p. 7).
Thörnqvist, T. (2015). "Sophistication, News and Individual Investor Trading". Mimeo. mimeo (cit. on p. 8).
Torgler, B. and M. Piatti (2013). *A Century of American Economic Review*. New York, New York: Palgrave Macmillan (cit. on p. 84).
Velte, P. (2018). "Does gender diversity in the audit committee influence key audit matters' readability in the audit report? UK evidence". *Corporate Social Responsibility and Environmental Management* 25 (5), pp. 748–755 (cit. on p. 86).
Wallace, L. S. et al. (2008). "Suitability and Readability of Consumer Medical Information Accompanying Prescription Medication Samples". *Patient Education and Counseling* 70 (3), pp. 420–425 (cit. on p. 8).
Weisberg, Y. J., C. G. De Young, and J. B. Hirsh (2011). "Gender Differences in Personality across the Ten Aspects of the Big Five". *Frontiers in Psychology* 2 (178) (cit. on p. 85).