UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Paper Economics

Cambridge Working Paper Economics: 1753

PUBLISHING WHILE FEMALE

ARE WOMEN HELD TO HIGHER STANDARDS? EVIDENCE FROM PEER REVIEW.

Erin Hengel

4 December 2017

I use readability scores to test if referees and/or editors apply higher standards to women's writing in academic peer review. I find: (i) female-authored papers are 1–6 percent better written than equivalent papers by men; (ii) the gap is two times higher in published articles than in earlier, draft versions of the same papers; (iii) women's writing gradually improves but men's does not—meaning the readability gap grows over authors' careers. In a dynamic model of an author's decision-making process, I show that tougher editorial standards and/or biased referee assignment are uniquely consistent with this pattern of choices. A conservative causal estimate derived from the model suggests senior female economists write at least 9 percent more clearly than they otherwise would. These findings indicate that higher standards burden women with an added time tax and probably contribute to academia's "Publishing Paradox" Consistent with this hypothesis, I find female-authored papers spend six months longer in peer review. More generally, tougher standards impose a quantity/quality tradeoff that characterises many instances of female output. They could resolve persistently lower—otherwise unexplained—female productivity in many high-skill occupations.

Publishing while Female

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

Erin Hengel†

November 2017

I use readability scores to test if referees and/or editors apply higher standards to women's writing in academic peer review. I find: (i) female-authored papers are 1–6 percent better written than equivalent papers by men; (ii) the gap is two times higher in published articles than in earlier, draft versions of the same papers; (iii) women's writing gradually improves but men's does not—meaning the readability gap grows over authors' careers. In a dynamic model of an author's decision-making process, I show that tougher editorial standards and/or biased referee assignment are uniquely consistent with this pattern of choices. A conservative causal estimate derived from the model suggests senior female economists write at least 9 percent more clearly than they otherwise would. These findings indicate that higher standards burden women with an added time tax and probably contribute to academia's "Publishing Paradox". Consistent with this hypothesis, I find female-authored papers spend six months longer in peer review. More generally, tougher standards impose a quantity/quality tradeoff that characterises many instances of female output. They could resolve persistently lower—otherwise unexplained—female productivity in many high-skill occupations.

1 Introduction

Ladies, we aren't that common in economics. Only a third, fifth and tenth of assistant, associate and full professors, respectively, are women (Romero, 2013). Female economists are less likely to make tenure, take longer when they do and earn much less than their male peers (Bandiera, 2016; Ceci et al., 2014; Ginther and Kahn, 2004; Weisshaar, 2017).[1]

These statistics are uncomfortable, but their causes are myriad: lower publishing rates, career choices, motherhood and, probably, bias. In lab 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;

*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 likewise indebted to Jeremy Edwards and my examination committee (Leonardo Felli and Hamish Low) for considerable input and advice. I also thank Oriana Bandiera, Cheryl Carleton, Gary Cook, Harris Dellas, Carola Frege, Jane Hunt, Brendan McCabe, Reshef Meir, Imran Rasul, Kevin Schnepel, Jarrod Zhang and participants at the Econometric Society European Winter Meeting, the Eastern Economic Association Conference, the Royal Economic Society Annual Conference and the European Meeting of the Econometric Society for useful comments. Finally, this paper could not have been written without substantial careful research assistance by Michael Hengel (my dad) and Eileen Hengel (my sister). All errors, of course, are mine.

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

[1]Weisshaar (2017) evaluates the probability of making tenure in Sociology, Computer Science and English departments. She finds female academics' lower productivity contributes to—but does not fully explain—tenure gaps in those fields.

Paludi and Bauer, 1983); when collaborating with men, women are given less credit (Heilman and Haynes, 2005; Sarsons, 2016).

This paper uses five reliable measures of writing clarity to show that women are likewise held to higher standards in peer review. (i) Female-authored articles published in top economics journals are better written than similar papers by men. The difference cannot be explained by year, journal, editor, topic, institution, English language ability or with various proxies for article quality and author productivity. (ii) The gap widens precisely while papers are being reviewed. I compare published articles to their pre-reviewed drafts. Forty percent of the gap originates *during* peer review. (iii) Female economists improve their writing; male economists don't. A dynamic model of an author's decision-making process shows that tougher editorial standards and/or biased referee assignment are the only explanations consistent with this pattern of choices. Using a conservative measure derived from the model, I estimate that this type of discrimination causes senior female economists to write at least 9 percent more clearly than they otherwise would. Finally, (iv) higher standards mean longer review—and as anticipated, female-authored papers take six months longer to complete peer review. This estimate is based on submit-accept times at *Econometrica*, and controls for, *inter alia*, motherhood, childbirth, citations and field.

Higher standards impose a quantity/quality tradeoff that likely contributes to academia's "Publishing Paradox" and "Leaky Pipeline".[2] Spending more time revising old research means spending less time conducting new research. Fewer papers results in fewer promotions, possibly driving women into fairer fields. Moreover, evidence of this tradeoff is present in a variety of occupations—*e.g.*, doctors, real estate agents and airline pilots—suggesting higher standards distort women's productivity, more generally.

Prior research indicates journal acceptance rates are genuinely bias-free (see, *e.g.*, Blank, 1991; Borsuk et al., 2009; Gilbert et al., 1994).[3] To the best of my knowledge, however, gender neutrality is established in only a narrow context (publication outcomes) using this single indicator. I ask a different question. Men's and women's papers may be published at comparable *rates*, but are they reviewed with comparable *scrutiny*? For, 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 well-intentioned referees might (unknowingly) inspect their papers more closely, demand a larger number of revisions and, in general, be less tolerant of complicated, dense writing.

Complicated, dense writing is my focus. In the English language, more 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).[4]

[2] "Publishing Paradox" and "Leaky Pipeline" refer to phenomena in academia whereby women publish fewer papers and disproportionately leave the profession, respectively.

[3] A possible exception is *Behavioral Ecology*, which increased its number of female first-authored papers after switching to double-blind review in 2001 (Budden et al., 2008a). Whether that increase was due to bias or the universal upward trend in female authorship, however, has been somewhat controversial (Budden et al., 2008b; Budden et al., 2008c; Webb et al., 2008; Whittaker, 2008).

[4] 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

If referees hold female- and male-authored papers to identical standards, both should be equally well written. To test this, I rely on a relationship familiar to linguists and educators: simple vocabulary and short sentences are easier to understand and straightforward to quantify. Using the five most widely used, studied and reliable formulas to exploit this, I analyse 9,123 article abstracts[5] published in the *American Economic Review* (*AER*), *Econometrica* (*ECA*), *Journal of Political Economy* (*JPE*) and *Quarterly Journal of Economics* (*QJE*).[6]

Female-authored abstracts are 1–6 percent more readable than those by men. Women write better despite controls for editor, journal, year and primary and tertiary *JEL* classification; that remains unchanged when proxying for article and author quality or accounting for English fluency. This means the readability gap probably wasn't (i) a response to specific policies in earlier eras; (ii) caused by women writing on topics that are easier to explain; (iii) due to a lopsided concentration of (non-)native English speakers;[7] nor (iv) generated by factors correlated with gender but really related to knowledge, intelligence and creativity.

Additionally, the gender readability gap widens *during* peer review. I compare National Bureau of Economic Research (NBER) working papers to their final, published versions; the gap is almost twice as large for the latter.[8] While both papers are exposed to many factors that impact readability, only published articles are subject to peer review. By comparing the two, influences unrelated to immediate peer review are isolated from those that are; assuming the former are not correlated with the latter's timing, a widening gap suggests a causal link.

Revising, redrafting and selecting just the right word is hard work; making sentences even marginally more readable takes time. Consistent with this hypothesis, I find female-authored papers spend ***six months longer*** in peer review. This estimate is based on submit-accept times from *Econometrica*, persists across a range of specifications and, in addition to other factors, controls for motherhood, childbirth, citations and field.[9]

Two explanations could account for these findings: either women voluntarily write better papers—*e.g.*, because they're more sensitive to referee criticism or overestimate the importance of writing well—or better written papers are women's response to external thresholds they do not control. Both imply women spend too much time rewriting old papers and not enough time writing new papers—but my evidence suggests the latter is primarily to blame.

In a dynamic model of an author's decision making process, I show that if women improve their writing over time and are not commensurately rewarded with higher acceptance rates (relative to men), then a persistent readability gap between equivalent peers is caused by discrimination. Authors improve readability only if they believe better writing leads to higher acceptance rates. And while oversensitivity and/or poor information may distort their

Shepherd, 1994, p. 172). In a random selection of 100 posts on Shit My Reviewers Say, a quarter deal with writing quality, document structure or word choice/tone.

[5] Readability scores are highly correlated across an article's abstract, introduction and discussion sections (Hartley et al., 2003a). See Section 2 for further discussion.

[6] For a discussion on the reliability of readability formulas, see DuBay (2004) and Section 2.1. A sixth commonly used measure is the Lexile Framework. Because its formula and software are proprietary, I do not include it in the analysis.

[7] It is not clear how—or even if—native English speakers write more clearly than non-native speakers. In fact, Hayden (2008) found that peer reviewed articles by the latter are more readable, on average.

[8] Many thanks to Kevin Schnepel for suggesting this idea.

[9] Predictably, giving birth slows down peer review. The coefficient on motherhood, however, is consistently negative (indicating a productivity boost) and almost always highly significant when subjected to several robustness checks. This result is provocative and discussed in Section 3.5. I encourage interpreting it with caution, however, given (i) counterintuitive results; (ii) the analysis did not intend to measure motherhood's impact on review times; and especially (iii) only a small number of mothers with young children are published in *Econometrica*.

beliefs—and affect readability—the impact declines with experience. Holding acceptance rates constant, this implies that a widening readability gap between equivalent authors is caused by discrimination—*i.e.*, asymmetric editorial standards and/or biased referee assignment beyond their control (Theorem 1).

Theorem 1 establishes sufficient conditions to demonstrate double standards are present in academic publishing: (i) experienced women write better than experienced men; (ii) women improve their writing over time; (iii) female-authored papers are accepted no more often than male-authored papers. Estimates from pooled subsamples at fixed publication counts suggest (i) and (ii) 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 probability of acceptance, conclusions from extensive study elsewhere are clear: "there are no sex differences in acceptance rates." (Ceci et al., 2014, p. 111; see also Section 3.4.3 for references to other research supporting this claim).

I also match prolific female authors to similarly productive male authors on characteristics that predict the topic, novelty and quality of research. In addition to explicitly accounting for author equivalence—the (principle) conditional independence assumption behind Theorem 1—matched pair comparisons: (i) identify the gender most likely to satisfy Theorem 1's conditions simultaneously;[10] and (ii) generate a (conservative) estimate of the effect of higher standards on authors' readability (Corollary 1).

Theorem 1's conditions were satisfied in 65 percent of matched pairs. In three quarters of those, the member discriminated against was female. Moreover, instances of obvious discrimination were predominately against women: the estimated effect of higher standards was almost twice as large in pairs suggesting discrimination against women; it clustered near zero for the small minority of pairs indicating discrimination against men. On average, higher standards cause senior female economists to write at least nine percent more clearly than they otherwise would.[11]

Asymmetric editorial standards and/or biased referee assignment affect women directly—as already discussed, women write more readably during and spend longer in peer review. They probably affect women's behaviour indirectly, too. As a final exercise, I compare papers pre- and post-review over increasing publication counts (Appendix A). In authors' earliest papers, the readability gap exclusively emerges during peer review; there is no gender difference in the draft readability of authors' first top publication. In later papers, women write well upfront; the gap chiefly materialises before peer review. This suggests that female economists adjust to higher standards *in* peer review by writing more clearly *before* peer review.

In economics, theoretical and empirical research on discrimination tends to focus on stereotype formation and belief structures motivating discriminatory actions (*e.g.*, Arrow, 1973; Becker, 1957; Bordalo et al., 2016; Coate and Loury, 1993; Phelps, 1972). This paper exclusively explores, in a non-laboratory environment, discrimination's impact on the behaviour and choices of people discriminated against.

[10] Each of Theorem 1's conditions must technically hold for the same author in two different situations—before and after gaining experience and when compared to an equivalent, experienced author of the opposite gender.

[11] While nine percent seems small, it is based on a single paragraph. Assuming a similar standard applies to every paragraph in a paper and improving each one takes slightly more time, the accumulated impact may be substantial. See also Berk et al. (2017) for a general discussion on how current culture may encourage extraneous (and time-consuming) demands in otherwise publishable papers.

This perspective has two advantages. First, it offers an alternative framework for studying the phenomenon. Discrimination is typically 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. Within a subjective expected utility framework, I show that authors' choices also reveal discrimination by editors and/or referees. Although context-specific, the model's basic logic—and its method of identifying discrimination—apply equally well to situations where people are repeatedly judged on and respond to feedback about some quantifiable component of their output.

Second, analysing discrimination from the perspective of people discriminated against forces us to think more deeply about its impact on, *inter alia*, occupational choice, worker motivation, human capital investment and, especially, productivity measurement. This paper joins a small, emerging literature examining these effects (*e.g.*, Craig and Fryer, 2017; Glover et al., 2017; Lavy and Sand, 2015; Parsons et al., 2011).[12]

Higher standards cause collateral damage to women's productivity. Unequal time spent making revisions leads to unequal time conducting new research; as a result, women write fewer papers.[13] Fewer papers justifies lower promotion rates.[14] If women seek fairer employment elsewhere—or quit the labour force entirely—it swells a "Leaky Pipeline".

Moreover, I also find evidence that female authors internalise tougher standards with strategies that disguise the underlying discrimination as voluntary choice. Women increasingly submit better written papers *ex ante* to offset biased evaluation *ex post*, meaning the readability gap between senior economists largely forms prior to—therefore appearing independent of—peer review. This pattern of behaviour obscures the line between personal preferences and external constraints and hints that academia overlooks other biases within its ranks.

Although analysed in a specific context—academia—higher standards impose a quantity vs. quality tradeoff that characterises many instances of female output. According to raw numerical counts, women produce less than men. Female reporters write fewer front-page bylines (Klos, 2014); female real estate agents list fewer homes (Seagraves and Gallimore, 2013); female physicians see fewer patients (Bloor et al., 2008)[15] and submit fewer grant proposals (Gordon et al., 2009); female pharmacists and lawyers work and bill fewer hours, respectively (Azmat and Ferrer, 2017; Goldin and Katz, 2016).

When ranked by narrowly defined outcome measures, however, women often outperform. Female students earn better grades (Voyer and Voyer, 2014); female auditors are more accurate and efficient (Chung and Monroe, 2001; Ittonen et al., 2013; Niskanen et al., 2011; O'Donnell and Johnson, 2001); congresswomen secure more federal funding for their

[12] A parallel research thread examines the broader impact of external signals (discriminatory or not) on women's behaviour (Kugler et al., 2017).

[13] A similar idea was also recently proposed in the philosophy literature (see Bright, 2017; Lee, 2016).

[14] Evidence on whether female academics are hired and promoted at lower rates is mixed. One study suggests so-called STEM (science, technology, engineering, mathematics) fields actually prefer hiring women—although male economists continue to show a slight (but not significant) preference for men (Williams et al., 2015). Other studies find male candidates are preferred in postdoctoral research and laboratory management positions (Moss-Racusin et al., 2012; Sheltzer and Smith, 2014). Men are also more likely granted tenure when compared to women with an identical publication history (Weisshaar, 2017) or for co-authored work (Sarsons, 2017). A study specific to the London School of Economics found female academics earn 12% less than men with identical experience and research productivity (Bandiera, 2016).

[15] Bloor et al. (2008)'s analysis considers only full-time (or maximum part-time), salaried physicians in the U.K. Similar results are found in Canada and the U.S., where physicians are paid on a per-service basis (Benedetti et al., 2004; Canadian Institute for Health Information, 2005).

districts, sponsor more legislation and score higher on a composite measure of legislative effectiveness (Anzia and Berry, 2011; Volden et al., 2013); houses listed by female real estate agents sell for higher prices (Salter et al., 2012; Seagraves and Gallimore, 2013);[16] patients treated by female physicians are less likely to die or be readmitted to hospital (Tsugawa et al., 2016); female pilots are involved in fewer fatal accidents (Bazargan and Guzhva, 2011; Vail and Ekman, 1986);[17] female economists write more clearly.

Additionally, if—like senior female economists—women internalise higher standards in somewhat roundabout ways, they could contribute to other labour market phenomena: sectoral and occupational concentration (Blau and Kahn, 2016; Cortés and Pan, 2016; Pertold-Gebicka et al., 2016); women's tendency to under negotiate pay (Babcock and Laschever, 2003)[18] and apply only to jobs they feel fully qualified for (Mohr, 2014). They may likewise reinforce work habits—*e.g.*, conscientiousness, tenacity and diligence—that correlate with quality and connote "femininity": female physicians consult longer with patients (Roter and Hall, 2004); female politicians fundraise more intensely (Jenkins, 2007);[19] female faculty commit fewer instances of academic misconduct (Fang et al., 2013); female lawyers make fewer ethical violations (Hatamyar and Simmons, 2004); female pharmacists are less likely to face performance-related disciplinary action (Schafheutle et al., 2011).[20]

Higher standards therefore offer another perspective to the gender gap in labour market outcomes. 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 stress inflexible working conditions (Goldin, 2014; Goldin and Katz, 2016), preferences (for a review, see, *e.g.*, Blau and Kahn, 2016) and policy design (Antecol et al., 2016). Still other research—which this paper joins—target more subtle forms of discrimination (*e.g.*, Sarsons, 2017; Wu, 2017). The gap probably emerges from all of these factors—and possibly many that are not yet identified. Equality means levelling the playing field for every one.

Furthermore, my results advocate using caution when employing shallow performance indicators in equations relating earnings (or other labour market outcomes) to gender. Higher standards raise quality at the expense of quantity. Performance indicators that weight the latter's fall more heavily than the former's rise will appear artificially low. If used to interpret gender wage gaps, they will undervalue women's work and confound estimates of labour mar-

[16] Seagraves and Gallimore (2013) find that normal houses (*i.e.* homes not sold under special sales conditions, such as foreclosures, fixer-uppers, corporate-owned properties, transfers and estate sales) sell at a significantly higher price when listed by a female real estate agent. The authors also find buyers pay less if they are represented by a male agent—although the effect is only present for homes sold under special sales conditions. An earlier study did not find any significant gender difference in selling performance for listing and selling agents (Turnbull and Dombrow, 2007).

[17] The evidence on general accident rates (including non-fatal accidents) is mixed. McFadden (1996) found no difference in female vs. male accident rates after adjusting for pilot experience and age. Walton and Politano (2016) found female accident rates were higher than male accident rates among inexperienced pilots but lower among experienced pilots.

[18] A more recent study suggests women do ask for higher pay—they just don't get it (Artz et al., 2016).

[19] Female politicians target a larger variety of potential donors using a wider array of methods (direct mail, television advertisements, *etc.*) (Jenkins, 2007).

[20] Evidence in several countries suggests female pharmacists are less likely to commit criminal offenses (prescription fraud, drug trafficking, etc.) and minor professional misdemeanours (inadequate written records, stock, etc.) (Payne and Dabney, 1997; Tullett et al., 2003). Self-reported survey evidence does not suggest female pharmacists make fewer dispensing errors (Szeinbach et al., 2007); evidence from a laboratory experiment indicates the opposite (Family et al., 2013). Similar gender trends have been found for physicians, dentists and other medical professionals (for a review of studies and discussion, see Firth-Cozens, 2008).

ket discrimination. A similar argument was recently made in a study of racial preferences in Major League Baseball. Parsons et al. (2011) find that race affects umpire calls, umpire calls influence players' behaviour and players' behaviour impacts performance metrics. As a result, common baseball statistics underestimate the talent of disadvantaged (usually minority) pitchers and overestimate the talent of advantaged (usually white) pitchers. An important contribution of my paper is to confirm this general point both in the context of gender discrimination and within a highly educated, professional working environment.[21]

This paper makes three final contributions. First, it adds to extensive (ongoing) research into peer review. Although mine, to the best of my knowledge, is the first to suggest and document evidence of gender bias in the peer review process (as opposed to its outcome), it joins contemporary or parallel research studying patterns in the editorial process (Card and DellaVigna, 2013; Clain and Leppel, 2017; Ellison, 2002) and bias in editorial decisions (Abrevaya and Hamermesh, 2012; Bransch and Kvasnicka, 2017; Card and DellaVigna, 2017).

Second, I use readability scores to untap a largely ignored, naturally occurring source of pseudo-experiments relevant to research on gender or racial bias—and differential group treatment, more generally.[22] Readability scores have their limitations (see Section 2.1) and their use in this manner applies to just a narrow set of questions. Nevertheless, they are cheaper than audit and correspondence studies and arguably more objective than survey data. An analogous approach may (or may not) expose similar group differences in, *inter alia*, successful business proposals funded by venture capitalists, letters to the editor published in newspapers or annual report introductions by CEOs.

Third, my findings emphasise the importance of transparency and monitoring. The least intrusive antidote to implicit bias is simple awareness and constant supervision. Unlike referee reports, journal acceptance rates are easy to measure and frequently audited; both factors foster accountability and encourage neutrality (Foschi, 1996). Monitoring referee reports is difficult, but it isn't impossible—especially if peer review were open. As discussed in Section 4, several science and medical journals not only reveal referees' identities, they also post reports online. Quality does not decline (it may actually increase), referees still referee (even those who initially refuse) and, given what's at stake, an extra 25–50 minutes spent reviewing seems tolerable (van Rooyen et al., 2010; van Rooyen et al., 1999; Walsh et al., 2000).

The remainder of the paper proceeds in the following order. Section 2 describes the data and readability measures. Analyses and results are presented in Section 3. They are succeeded by a detailed discussion (Section 4) and conclusions (Section 5).

[21] Another recent study might also illustrate this point. Glover et al. (2017) find that obvious productivity measures decline when minority grocery store workers are overseen by biased managers. If due to demotivation or inattention by managers—as the authors propose—their behaviour reinforces statistical discrimination. On the other hand, slower checkout times, less overtime work and seeing fewer customers could result from biased managers being more critical of minorities' work (*e.g.*, minority workers are more likely to be punished for an incorrect amount of money in the till, not immediately clocking out at the end of a shift or accidentally scanning a single item multiple times).

[22] Using readability scores to uncover gender bias in the way news is reported was first proposed by Ali et al. (2010). In an effort to determine gender differences in writing styles, Hartley et al. (2003b) compare male and female Flesch Reading Ease scores for 80 papers published in the *Journal of Educational Psychology*; they found no consistent, sex-specific difference. See Footnote 129 and Footnote 130 for a discussion and list of other creative ways readability scores have been used in academic research.

TABLE I: 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	695	520	408	413	2,036
2010–15	732	384	181	251	1,548
Total	2,083	3,118	2,446	1,475	9,122

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.

2 Data

The data include every English article published in *AER*, *Econometrica*, *JPE* and *QJE* between January 1950 and December 2015 (inclusive). Prior research has found authors write in a stylistically consistent manner across the abstract, introduction and discussion section of a peer reviewed article (Hartley et al., 2003b).[23] Of these three, I concentrate on abstracts. Abstract structure is standardised in a manner optimal for computing readability scores: 100–200 words, no citations and few abbreviations and equations (Dale and Chall, 1948). Abstracts are self-contained, universally summarise the research and are the first and most frequently read part of an article (King et al., 2006)—all factors suggesting a relatively homogenous degree of review across journals and subject matter. Conveniently, most have also been converted to accurate machine-readable text by digital libraries and bibliographic databases.

The largest sample comes 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.[24] Table I displays data coverage by journal and decade. Bibliographic information and PDFs were scraped from the websites of Oxford Journals, the American Economic Association, the Econometric Society, Wiley, JSTOR and EBSCO.

Based on authors' given names, gender was assigned via 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 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. In the remaining cases, I emailed or telephoned colleagues and institutions associated with the author.

[23] Within-manuscript correlations of Flesch Reading Ease scores are 0.64 (abstracts vs. introductions) to 0.74 (abstracts vs. discussions), suggesting "authors are remarkably consistent in how they use word categories" (Hartley et al., 2003a, p. 392).

[24] Unless otherwise mentioned, observations exclude the May issue of *AER* (*Papers & Proceedings*).

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. I create 64 dummy variables, each of which represents one or more institution(s); groupings reflect counts of distinct articles in which an institution was listed as an affiliation.[25] Specifically, institutions listed in 59 or fewer articles were grouped in bins of 10 to form six dummy variables: the 751 institutions mentioned in 0–9 articles were grouped to form the first dummy variable, the 92 mentioned in 10–19 articles were grouped to form the second, *etc.* Fifty-eight institutions were affiliated with 60 or more articles; each is assigned its own dummy variable. 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.

I control for article quality and author productivity in several ways. First, I use article citations from the Web of Science database. Second, I generate 30 dummy variables that group authors by career-total publication counts in the four journals. For example, Daron Acemoglu and Jean Tirole form one group (each published 45 articles as of December 2015); Alvin Roth, Elhanan Helpman and Gene Grossman form another (27 articles).[26] In Section 3.3 and Section 3.5, I additionally control for the number of prior top-four papers (at time of publication). For co-authored articles, only the data corresponding to the most prolific author is used.[27]

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;[28] 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).[29]

[25] 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.

[26] 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 45 articles, but only one author has published 37 (Andrei Shleifer).

[27] 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.

[28] 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).

[29] 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.

I create dummy variables corresponding to the 20 primary *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.

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.[30] Editor controls are based on distinct lead editor/editorial boards—*i.e.*, they differ by at least one member. In total, 74 groups are formed in this manner.

The analysis in Section 3.3 matches published articles with NBER working papers. Matches were first attempted using citation data from RePEc and then by searching 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.[31] Bibliographic information and abstract text were scraped from www.nber.org.

The analysis in Section 3.5 compiles submit-accept times at *Econometrica*—the only journal that makes any kind of disaggregated data on the revision process publicly available.[32] I extracted this information from digitised articles using the open source command utility `pdftotext`.

To control for motherhood's impact on revision times, I recorded children's birth years for women with at least one 100 percent 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, intelius.com background checks and local school district/popular extra-curricular activity websites.[33] 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. If an exhaustive search turned up no reference to children, I assumed the woman in question did not have any.[34]

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

[31] Because a small number of NBER working papers were eventually published as multiple articles or combined into a single paper, the mapping is not one-for-one.

[32] Printed at the end of every *Econometrica* article published on or after March 1970 that was not originally presented as an Econometric Society lecture is the date it was first submitted and the date final revisions were received. Before 1970, only "A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators" (January, 1965) included this information. "Separable Preferences, Strategyproofness, and Decomposability" (May, 1999) only printed the year of submission; I assume the month is January.

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

[34] In several instances, I obtained this information from acquaintances, friends and colleagues or by asking the woman directly. Given its sensitive nature, children's birth years are not currently available on my website (unlike other data in this paper).

TABLE II: Readability scores

Score	Formula
Flesch Reading Ease	$206.835 - 1.015 \times AWS - 84.6 \times ASW$
Flesch-Kincaid	$-15.59 + 0.390 \times AWS + 11.8 \times ASW$
Gunning Fog	$0.4 \times \left(AWS + 100 \times PWW\right)$
SMOG	$3.1291 + 5.7127 \times \sqrt{APS}$
Dale-Chall	$3.6365 + 0.0496 \times AWS + 15.79 \times DWW$

Notes. AWS: average number of words per sentence; ASW: average number of syllables per word; PWW: ratio of polysyllabic words (3+ syllables) to word count; APS: average number of polysyllabic words per sentence; DWW: ratio of difficult words (not on Dale-Chall list) to word count.

2.1 Measuring readability

Advanced vocabulary and complicated sentences are the two strongest predictors of readability (Chall and Dale, 1995; DuBay, 2004). Most readability formulas exploit this relationship, combining frequency of easy words with sentence length to arrive at a single score.

Although hundreds exist, I concentrate on the five most widely used, tested and reliable measures 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 Table II.

The Flesch Reading Ease scales from 0 (hard) to 100 (easy). In contrast, the other four scores generate grade levels estimating the minimum years of schooling necessary to confidently understand an evaluated text—and so lower scores indicate easier-to-read text. To minimise confusion, I multiply the four grade-level scores by negative one. Thus, higher numbers universally correspond to clearer writing throughout the paper.

The constants in each formula vary widely as do the components used to rank vocabulary. The Flesch Reading Ease and Flesch-Kincaid scales rely on syllable count, Gunning Fog and SMOG total polysyllabic words (words with three or more syllables) while Dale-Chall tallies words not on a pre-defined list of 3,000 so-called "easy" words.[35] These differences mean the four grade-level scores rarely generate identical figures; nevertheless, all five scores produce roughly equivalent rankings (Begeny and Greene, 2014).

Criticisms of readability scores are usually levied at their imprecision.[36] Evidence suggests they may not be accurate enough to adequately assess or guide development of legal briefs (Sirico, 2007), financial disclosure documents (Loughran and McDonald, 2014) or school reading material (Ardoin et al., 2005; Powell-Smith and Bradley-Klug, 2001). But despite poor accuracy, readability scores *do* correlate with reading difficulty (Begeny and Greene, 2014; DuBay, 2004; Francis et al., 2008; Hintze and Christ, 2004) making them appropriate measures to estimate gender differences in large samples.[37]

[35] Specifically, 3,000 words understood by 80 percent of fourth-grade readers (aged 9–10).

[36] Another criticism of readability formulas is that their use encourages writers to shorten sentences and chose simpler vocabulary at the expense of comprehension (for a discussion, see DuBay, 2004; Long and Christensen, 2011). This study implicitly assumes that the authors of papers published in the four journals and time periods covered by the data have not "written to the formula" in any meaningful (or gender-specific) way.

[37] At a bare minimum, no study (to my knowledge) has ever shown that any of the five scores used here are significantly inversely related to reading difficulty. Evidence from Begeny and Greene (2014) suggests the four grade-level readability scores, and particularly the SMOG and Dale-Chall scores, are more accurate for higher ability readers. (The study did not assess the Flesch Reading Ease score.)

A second criticism of readability scores is practical. Some programs that calculate them rely on unclear, inconsistent and possibly inaccurate algorithms to count words, sentences and syllables and determine whether a word is on Dale-Chall's easy word list (for a discussion, see Sirico, 2007). Additionally, features of the text—particularly full stops used in abbreviations and decimals in numbers—frequently underestimate average words per sentence and syllables per word.[38]

To transparently handle these issues and eliminate ambiguity in how the readability scores were calculated, I wrote the Python module `Textatistic`. Its code and detailed documentation is available at GitHub. A brief description is provided here.

To determine sentence count, the program replaces common abbreviations with their full text,[39] decimals with a zero and deletes question and exclamation marks used in an obvious, mid-sentence rhetorical manner.[40] 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.[41]

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.

3 Analyses and results

Analyses and results are organised as follows. In Section 3.1, I scrutinise readability at the article level, controlling for editor, journal, year, journal and year interactions, institution, author productivity, article quality (citation count), English fluency and field. The results suggest a gap does indeed exist. They also rule out obvious confounding factors—women writing on easier topics, editorial policies in earlier eras, *etc.* Next (Section 3.2), I investigate readability at the author-level in a fixed effects regression. This accounts for author-specific productivity, quality and other effects that influence writing—*e.g.*, innate talent—but are otherwise unconnected to peer review.

[38] Typesetting code used to render equations—common in *Econometrica* abstracts published before 1980—also affects the accuracy of readability scores. I therefore manually replaced all such code with equivalent unicode characters. 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. Readability scores were determined using the modified text.

[39] 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.

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

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

TABLE III: Textual characteristics per sentence, by gender

	Men	Women	Difference
No. characters	134.73	130.27	4.46***
	(0.43)	(1.45)	(1.57)
No. words	24.16	23.06	1.10***
	(0.08)	(0.27)	(0.29)
No. syllables	40.65	38.65	2.01***
	(0.13)	(0.45)	(0.48)
No. polysyllabic words	4.69	4.31	0.39***
	(0.02)	(0.07)	(0.08)
No. difficult words	9.38	8.91	0.48***
	(0.03)	(0.12)	(0.13)

Notes. Sample 9,122. Figures from an OLS regression of female ratio on each characteristic divided by sentence count. Male effects estimated at a ratio of zero; female effects estimated at a ratio of one. Robust standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

In Section 3.3, I match published articles—which have gone through peer review—to earlier, draft versions of the same papers—which have not. Assuming timing independence, this isolates the effect of peer review and causally links it to the gender readability gap. Section 3.4 takes the final step and causally links the gap to referees and/or editors. I first develop a dynamic model of an author's decision-making process to evaluate the remaining alternatives (Section 3.4.1): gender differences in biology/behaviour and/or knowledge about referee expectations. Based on the model, I propose a method for identifying the impact of discrimination on authors' readability. In Section 3.4.3, I use matching to estimate it.

Finally, prolonged peer review should be one observable repercussion from subjecting female authors to higher standards. Using submit-accept times from *Econometrica*, I evaluate this hypothesis, controlling for, *inter alia*, motherhood, childbirth, citations and field (Section 3.5).

3.1 Article-level analysis

Table III displays each gender's average per sentence number of characters, words, syllables, polysyllabic words and difficult words. Women write shorter, simpler sentences—they contain fewer characters, fewer syllables, fewer words and fewer "hard" words. Differences are highly statistically significant.

Table IV presents coefficients from an ordinary least squares (OLS) regression of the ratio of female co-authors on the five readability scores. To account for error correlation by editorial policy, observations are grouped by journal editor/editorial board and standard errors are adjusted accordingly.[42]

Column (1) controls for journal and editor: abstracts written only by women score about one point higher on the Flesch Reading Ease scale; according to the four grade-level measures, they take 1–6 fewer months of schooling to understand.[43] Percentage-wise, women

[42] Standard errors are very similar when clustering at the volume-, issue- or paper-level (see Hengel, 2016, p. 39–41).

[43] Coefficients from regressions on Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall scores represent the marginal effect in years of schooling. Monthly figures found by multiplying each coefficient by 12.

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Flesch Reading Ease	0.90*	0.87*	0.83*	0.81	0.97*	0.52	0.92
	(0.48)	(0.48)	(0.50)	(0.48)	(0.50)	(0.53)	(0.71)
Flesch-Kincaid	0.19*	0.18	0.18	0.19*	0.22*	0.23*	0.25*
	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.12)	(0.14)
Gunning Fog	0.33***	0.33***	0.33***	0.33***	0.37***	0.34**	0.36**
	(0.12)	(0.12)	(0.12)	(0.13)	(0.14)	(0.14)	(0.16)
SMOG	0.21**	0.21**	0.22**	0.21**	0.23**	0.19*	0.23*
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)	(0.12)
Dale-Chall	0.10**	0.10**	0.10**	0.09**	0.11**	0.09*	0.13**
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓
JEL (primary) effects						✓	
JEL (tertiary) effects							✓

Notes. 9,122 articles in (1)–(5); 5,216 articles in (6); 5,777 articles—including 561 from *AER Papers & Proceedings* (see Footnote 46)—in (7). Figures represent the coefficient on female ratio from an OLS regression on the relevant readability score. Quality controls denoted by ✓[1] include citation count and max. T_j fixed effects. Standard errors clustered on editor in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

write 1–2 percent better than men.[44]

Column (2) includes 63 year dummies; column (3) adds another 182 journal and year interaction dummies; columns (4) and (5) introduce 64 institution effects, quality controls—citation count and 30 max. T_j effects (maximum co-author lifetime publication count for paper j)—and a dummy variable capturing English fluency.[45] Coefficients and standard errors in columns (2)–(5) are very similar to those in column (1).

The coefficients on the journal dummies in (2) are presented in Appendix D.1. They compare *AER*'s readability to the readability of *Econometrica*, *JPE* and *QJE*, providing a useful check on the reliability of readability formulas in the context of economic writing. As intuitively expected, all five scores agree that *Econometrica* is harder to read; four out of five scores suggest *JPE* is, too, while *QJE* is easier.

Columns (6) and (7) control for primary *JEL* classification. (6) includes 20 fixed effects for primary *JEL* categories; (7) includes 718 effects for tertiary categories. Due to small sample sizes, (7) includes 561 articles from *AER Papers & Proceedings*.[46] Since only post–

[44] Quotient of the coefficient on female ratio divided by the effect for men (ratio of zero) estimated at other co-variates' observed values (see Appendix D.1).

[45] In Hengel (2016, p. 44 and p. 46), I include controls for the order an article appears in an issue—another measure of a paper's quality. Results are similar to those in Table IV. In addition to the control from English fluency presented here, see Hengel (2016, pp. 35–36) for further evicence that the female authors in my data are no more or less likely to be native English speakers.

[46] *AER Papers & Proceedings* is coded as a separate journal and edited by the American Economic Association's president-elect. *AER Papers & Proceedings* does not publish abstracts in its print version; only select years and papers are available online (2003 and 2011–2015), all of which are included. Excluding these articles does not

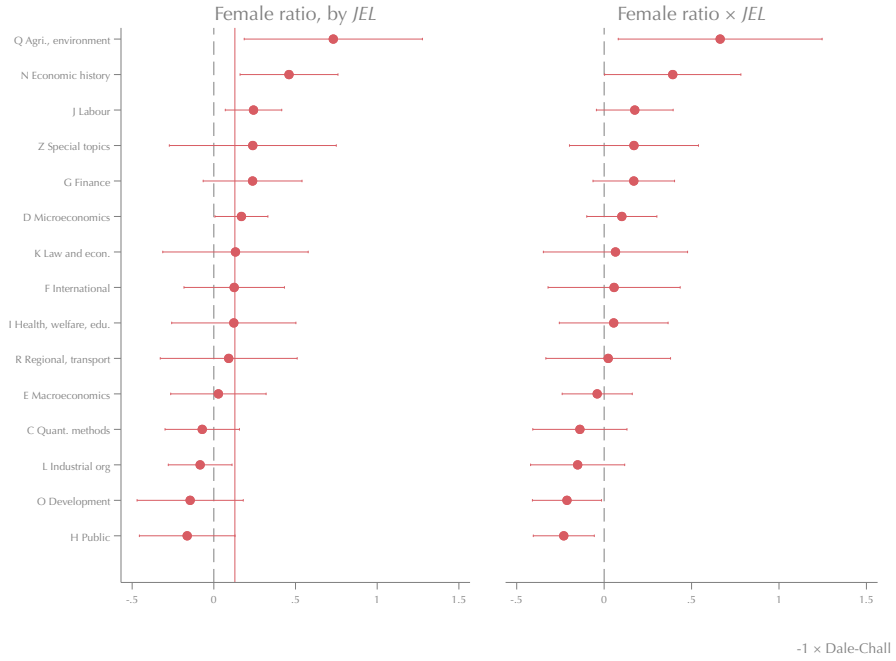


FIGURE I: Gender differences in readability, by *JEL* classification

Notes. Sample 5,777 articles, including 561 from *AER Papers & Proceedings* (see Footnote 46). Codes A, B, M and P dropped due to small sample sizes of female-authored papers (see Footnote 47). 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 15×1 column vector with kth entry a binary variable equal to one if article j is classified as the kth *JEL* code; $\mathbf{X}_j$ is a vector of editor, journal, year, institution, English language dummies and quality controls (citation count and max. T_j fixed effects); ε_j is the error term. Left-hand graph shows marginal effects of female ratio for each *JEL* code ($\beta_1 + \beta_3^k$); the pink vertical line is the mean effect at observed *JEL* codes (0.128, standard error 0.046). Right-hand graph displays interaction terms (β_3^k). Horizontal lines represent 90 percent confidence intervals from standard errors adjusted for clustering on editor.

1990 *JEL* classifications are used, estimates in both columns exclude over 40 percent of the data. Nevertheless, coefficients and standard errors are roughly equivalent.

Figure I 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 and max. T_j fixed effects.[47] Again, due to small samples—particularly of female authors—Figure I includes 561 articles from *AER Papers & Proceedings*.[48]

The pink vertical line in Figure I's left-hand graph is the marginal effect of female authorship at the mean. Its estimate coincides with results in Table IV—women's papers require six fewer weeks of schooling to understand—and is highly significant. Points reflect marginal effects across *JEL* classification; bars represent 90 percent confidence intervals from standard errors clustered by editor.

Women earn higher marks for clarity in 11 out of 15 categories; only three are at least

impact results or conclusions—coefficients are almost identical to those in column (6), but standard errors are somewhat higher. (Analysis not shown, but is available on request: erin.hengel@gmail.com.)

[47] 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.

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

weakly significant: Q (Agricultural and Natural Resource Economics; Environmental and Ecological Economics), N (Economic History), and J (Labour Economics). 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 I's right-hand graph displays coefficients from interacting the ratio of female co-authors with each *JEL* code. Q and N are significantly above the mean, O and 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 I 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. 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.

In the remainder of the paper, I do not explicitly control for *JEL* classification (unless otherwise specified). Comparable codes are available for only a subset of the data and Table IV and Figure I suggest they are relatively unimportant, anyway.

3.2 Author-level analysis

I next 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 kth author. I then estimate the dynamic panel model in Equation (1):

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

$R_{j_{it}}$ is the readability score for article j—author i's tth publication; R_{it-1} is the corresponding value of author i's $t-1$th paper. Gender enters twice—the binary variable male_i and 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 and English fluency dummies; quality controls—citation count and max. T_j fixed effects; and N_j to account for author i's proportional contribution to paper j. α_i are author-specific effects and ε_{it} is an idiosyncratic error. α_i are eliminated by first-differencing; endogeneity in the lagged dependant variable is instrumented with earlier lags (Arellano and Bover, 1995; Blundell and Bond, 1998). To account for duplicate articles, the regression is weighted by $1/N_j$.[49] Standard errors are adjusted for two-way clustering on editor and author.

Table V displays results. Rows one and two present contemporaneous marginal effects on co-authoring with women for female (β_1) and male ($\beta_1 + \beta_2$) authors, respectively. Both estimates are positive—everyone writes more clearly when collaborating with women. Marginal effects for women are highly significant and at least twice as large as those in

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

TABLE V: Gender differences in readability, author-level analysis

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Female ratio (women)	2.37**	0.35*	0.66***	0.48**	0.23**
	(1.00)	(0.20)	(0.24)	(0.19)	(0.10)
Female ratio (men)	0.57	0.09	0.15	0.09	0.10
	(1.31)	(0.25)	(0.30)	(0.21)	(0.11)
Female ratio×male	−1.80	−0.26	−0.50	−0.38	−0.14
	(1.53)	(0.32)	(0.37)	(0.27)	(0.13)
Lagged score	0.03**	0.04***	0.03*	0.03*	0.03**
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)
z-test for no serial correlation					
Order 1	−20.26	−15.97	−17.14	−19.93	−20.75
Order 2	0.56	−0.22	0.08	0.19	−0.73
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,186 observations (2,827 authors). Figures from first-differenced, IV estimation of Equation (1) (Arellano and Bover, 1995; Blundell and Bond, 1998). Female ratio (women): contemporaneous marginal effect of a paper's female co-author ratio for female authors (β_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 ✓[1] include citation count and max. T_j fixed effects. Regressions weighted by $1/N_j$; standard errors adjusted for two-way clustering on editor and author (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

Table IV—women write 2–6 percent better than men.[50] When men write with women, however, marginal effects are smaller and less precise.

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 V suggests they don't. While the interaction terms (β_2) are insignificant—*i.e.*, the observed disparity is plausibly due to chance—the difference may reveal an increasing, convex relationship between female ratio and readability. Men's smaller effect potentially reflects their disproportionate tendency to co-author exclusively with other men—precisely where the marginal impact of an additional woman is low.[51]

Tests for serial correlation indicate no model misspecification. Coefficients on the lagged dependant variables are small, suggesting readability is mostly determined contemporaneously, possibly during the revision process. Nevertheless, their uniform positivity and significance indicate modest persistence.

[50] Quotient of β_1 divided by the total effect for men co-authoring with no women (female ratio of zero) estimated at other co-variates' observed values (see Appendix D.2).

[51] 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).

3.3 Comparing abstracts pre- and post-review.

Table IV establishes a gender readability gap for abstracts published in top economics journals. Table V suggests it primarily forms contemporaneously. A possible contemporaneous cause is peer review—specifically referee and/or editor demands for more revisions by female authors.

In this section, I show that peer review does indeed cause (or exacerbate) the gender readability gap. To do so, I analyse papers before and after review by comparing published articles to their draft versions. Assuming peer review is the sole gender-related factor to affect abstract readability between versions, a larger increase in women's readability relative to men's is evidence of causality.

3.3.1 Summary statistics. As discussed in Section 2, drafts were collected from NBER Technical and Working Paper Series. NBER series were used as the exclusive data source for two reasons. First, approximately one-fifth of articles in the data were originally part of an NBER series, making it the largest single source of draft papers. Second, NBER persistently releases its working papers two to three years before publication (mean 2.1 years)—precisely the length of time spent in peer review (Ellison, 2002; Goldberg, 2015).

Table VI compares textual characteristics between versions. Means in the first three columns are of majority male-authored papers (female ratio strictly below 50 percent); the final three columns are majority female-authored papers (female ratio at or above 50 percent).

Abstracts are considerably altered during peer review. Table VI's first panel displays raw counts. 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.

Peer review's impact on readability, however, is unclear. Readability scores are weighted averages of the ratios of (i) total word or "hard" word to sentence count and (ii) hard word to word count. Between working paper and published versions, (i) decreases and (ii) increases (Table VI, second panel).[52] (i) Peer review shortens sentences and reduces hard words per sentence: in male-authored papers, sentences are 5 percent shorter and contain 26 percent fewer polysyllabic words; in female-authored papers, they are 7 percent shorter and contain 30 percent fewer polysyllabic words. (ii) As a fraction of total word count, however, syllables, polysyllabic words and difficult words rise. To wit, hard word counts and total word count decline, but the latter by proportionately more; their ratios increase: between 1–3 percent for men and 1–2 percent for women.

According to the majority of scores, peer review improves readability (Table VI, third panel), a finding consistent with similar investigations at medical journals (Biddle and Aker, 1996; Hayden, 2008; Roberts and Nolen-Hoeksema, 1994).[53] Thanks to fewer hard words

[52] A greater decline in total word count relative to hard word count may be specific to abstracts, which are edited for length as well as readability. In an analysis of abstracts, introductions and discussions, abstract sentences were shorter but contained more hard words; overall, they had the lowest Flesch Reading Ease scores (Hartley et al., 2003a).

[53] Hayden (2008) found no significant change in the Flesch Reading Ease score during peer review itself (submission vs. acceptance), but a significant positive effect from post-acceptance editing by the journal editor and a copy-editor. Compared to economics journals, however, medical journals ask for fewer revisions (Ellison, 2002; Hayden, 2008) and enjoy substantially shorter review times (see, *e.g.*, Journal of Trauma and Acute Care

Table VI: Textual characteristics, published papers vs. drafts

	Men			Women		
	Working paper	Published article	Difference	Working paper	Published article	Difference
No. sentences	6.47	5.10	−1.375***	6.77	5.06	−1.711***
	(0.06)	(0.04)	(0.054)	(0.15)	(0.08)	(0.139)
No. characters	862.45	649.68	−212.767***	907.36	635.97	−271.385***
	(7.19)	(4.67)	(7.160)	(18.53)	(10.31)	(18.439)
No. words	155.70	115.70	−40.004***	164.45	113.63	−50.813***
	(1.32)	(0.85)	(1.323)	(3.42)	(1.91)	(3.428)
No. syllables	257.01	193.36	−63.653***	269.02	187.78	−81.242***
	(2.15)	(1.40)	(2.135)	(5.54)	(3.08)	(5.504)
No. polysyllabic words	28.36	21.81	−6.545***	28.93	20.63	−8.308***
	(0.28)	(0.18)	(0.245)	(0.71)	(0.41)	(0.627)
No. difficult words	58.51	44.61	−13.892***	60.32	42.37	−17.949***
	(0.51)	(0.33)	(0.482)	(1.30)	(0.74)	(1.204)
No. words / sentence count	24.74	23.58	−1.166***	24.98	23.16	−1.820***
	(0.14)	(0.12)	(0.124)	(0.33)	(0.27)	(0.302)
No. polysyllabic words / sentence count	6.03	4.45	−1.576***	6.05	4.23	−1.819***
	(0.07)	(0.03)	(0.060)	(0.18)	(0.08)	(0.155)
No. syllables / word count	1.66	1.68	0.018***	1.64	1.66	0.015***
	(0.00)	(0.00)	(0.002)	(0.01)	(0.00)	(0.004)
No. polysyllabic words / word count	0.18	0.19	0.006***	0.18	0.18	0.005**
	(0.00)	(0.00)	(0.001)	(0.00)	(0.00)	(0.002)
No. difficult words / word count	0.38	0.39	0.009***	0.37	0.37	0.006**
	(0.00)	(0.00)	(0.001)	(0.00)	(0.00)	(0.002)
Flesch Reading Ease	41.46	41.13	−0.332*	42.51	43.08	0.564
	(0.26)	(0.18)	(0.185)	(0.66)	(0.43)	(0.452)
Flesch-Kincaid	−13.62	−13.38	0.243***	−13.53	−13.00	0.531***
	(0.06)	(0.05)	(0.050)	(0.15)	(0.11)	(0.122)
Gunning Fog	−17.28	−17.04	0.242***	−17.13	−16.58	0.547***
	(0.07)	(0.05)	(0.055)	(0.18)	(0.13)	(0.140)
SMOG	−15.14	−15.00	0.135***	−15.02	−14.70	0.327***
	(0.05)	(0.03)	(0.035)	(0.13)	(0.09)	(0.095)
Dale-Chall	−10.85	−10.93	−0.084***	−10.71	−10.70	0.003
	(0.02)	(0.02)	(0.016)	(0.06)	(0.04)	(0.037)

Notes. Sample 1,714 published articles authored by more than 50 percent men (1,715 NBER working papers); 364 published articles authored by at least 50 percent women (365 NBER working papers). Figures are means of textual characteristics by sex for NBER working papers and published articles. Third and sixth columns subtract working paper figures (columns 1 and 4) from published article figures (columns 2 and 5) for men and women. Standard errors in parentheses. ***, ** and * difference statistically significant at 1%, 5% and 10%, respectively.

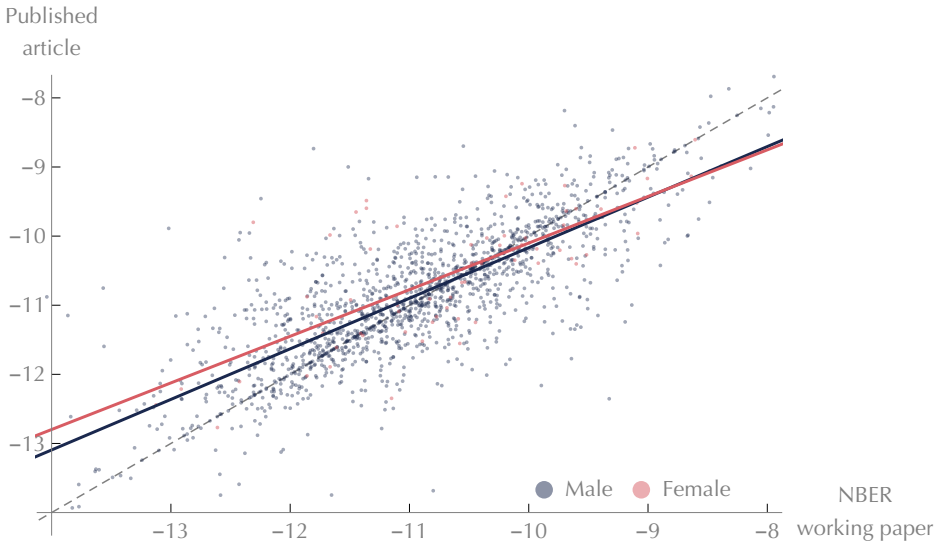


FIGURE II: Published paper vs. draft readability

Notes. Sample 1,631 NBER working papers; 1,629 published articles. Data points represent each abstract's $-1 \times$ Dale-Chall score pre-publicaction (NBER working paper) plotted against its $-1 \times$ Dale-Chall post-publication score. Pink represents women co-authoring only with other women (65 NBER working papers; 64 published articles); blue are men co-authoring only with other men (1,566 NBER working papers; 1,565 published articles); articles co-authored by men and women are omitted. The line of best fit using OLS is shown separately for men and women. The grey dashed line is the 45 degree line through the origin; points above (below) it denote abstracts that were better written after (before) peer review.

per sentence, SMOG scores are higher in published articles regardless of gender (see Table II). In female-authored papers, the net effect for remaining scores is similarly positive. In male-authored papers, however, only the Gunning Fog and Flesch-Kincaid scores indicate a positive net effect; for the Flesch Reading Ease and Dale-Chall scores, it's negative. In any case, women's papers endure comparatively greater cuts in hard words relative to total words and larger falls in words per sentence; their abstracts always become more readable during peer review than do those by men.

Figure II reiterates women's readability gains. It plots draft Dale-Chall scores (x-axis) against abstracts' published scores (y axis) for men (blue) and women (pink). The grey, dashed line is a 45 degree line through the origin. As might be expected, poorly written draft abstracts emerge more readable in the published version (above the 45 degree line); abstracts that were already well written come out slightly less so (below the 45 degree line). Regardless, female-authored published papers are again more readable than they were as working papers relative to male-authored papers—further evidence that women's papers are more heavily scrutinised during peer review.[54]

3.3.2 Identification. The data 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 gender's contemporaneous effect on peer review scrutiny. The key equation connects published articles to earlier versions of the same paper: scores depend on draft readability

Editorial Board, 2015), suggesting pre-acceptance readability edits are less common.

[54] An alternative hypothesis consistent with Figure II is that male-authored papers are scrutinised more, but edits made as a result reduce readability. The more substantial changes made to female-authored papers documented in Table VI, however, contradicts this theory.

as well as factors that affect writing clarity any time *after* being released as working papers. Equation (2) is the OLS representation of this relationship.

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

where R_{jP} and R_{jW} are readability scores for working (W) and published (P) versions of paper j, respectively. β_{0P} is a constant specific to version P; β_{1P} is the coefficient of interest and reflects the particular impact female ratio$_j$ has in peer review. $\mathbf{X}_{jP}$ and μ_{jP} are P-specific observable (editor, journal, journal-year interactions and English language dummies and max. t_j) and unobservable components, respectively.[55] ε_{jP} is P's error term.

P-specific variables may be correlated with R_{jW}. Even if μ_{jP} and female ratio$_j$ remain independent, positive correlation between R_{jW} and female ratio$_j$ (Table VI) still biases OLS estimates of β_{1P} in a direction opposite to the bias on R_{jW}. Equation (3) eliminates the distortion by subtracting R_{jW} from both sides of Equation (2):

$$R_{jP} - R_{jW} = \beta_{0P} + \beta_{1P}\,\text{female ratio}_j + \boldsymbol{\theta}_P \mathbf{X}_{jP} + \mu_{jP} + \varepsilon_{jP}. \tag{3}$$

Assuming zero partial correlation between female ratio$_j$ and μ_{jP}, OLS generates an unbiased estimate of β_{1P}.

An alternative strategy based on Ashenfelter and Krueger (1994) separately estimates NBER working paper and published article readability using generalised least squares (GLS); β_{1P} is identified post-estimation by differencing coefficients. The set-up combines Equation (2) with a relationship defining readability scores *before* external evaluators demand edits (Equation (4)).

$$R_{jW} = \beta_{0W} + \beta_{1W}\,\text{female ratio}_j + \boldsymbol{\theta}_W \mathbf{X}_{jW} + \mu_{jW} + \varepsilon_{jW}, \tag{4}$$

where β_{0W} is a constant specific to version W and β_{1W} reflects female ratio$_j$'s impact on readability prior to peer review. $\mathbf{X}_{jW}$ and μ_{jW} are version-invariant observable (publication year, citation count and max. T_j) and unobservable components, respectively.[56] ε_{jW} is version W's error term.

OLS estimates of Equation (4) may be biased by arbitrary correlation between μ_{jW} and the explanatory variables. Equation (5) defines a general structure for that correlation (Ashenfelter and Krueger, 1994).

$$\mu_{jW} = \gamma + \eta\,\text{female ratio}_j + \boldsymbol{\delta}_W \mathbf{X}_{jW} + \boldsymbol{\delta}_P \mathbf{X}_{jP} + \omega_j, \tag{5}$$

where ω_j is uncorrelated with female ratio$_j$, $\mathbf{X}_{jW}$ and $\mathbf{X}_{jP}$. Substituting Equation (5) into Equation (4) generates the following reduced form representation of R_{jW}:

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

[55] max. t_j is the number of prior papers published in any of the top four economics journals by article j's most prolific co-author. It and the English language dummy are considered P-specific because they may influence the degree to which editors and/or referees scrutinise the paper. Because all papers in both samples share the same highest-ranked institution (NBER), authors' institutions—which presumably have a similar effect—are omitted.

[56] I assume the duration between a paper's NBER release and its publication is too short to influence aggregate time trends; publication year dummies are applied to both working paper and published versions.

where $\widetilde{\beta}_{0W} = \beta_{0W} + \gamma$, $\widetilde{\beta}_{1W} = \beta_{1W} + \eta$, $\widetilde{\boldsymbol{\theta}}_W = \boldsymbol{\theta}_W + \boldsymbol{\delta}_W$ and $\widetilde{\varepsilon}_{jW} = \varepsilon_{jW} + \omega_j$. Similarly, obtain R_{jP}'s reduced form by substituting Equation (6) into Equation (2):

$$\begin{aligned} R_{jP} = (\widetilde{\beta}_{0W} + \beta_{0P}) + (\widetilde{\beta}_{1W} + \beta_{1P})\,\text{female ratio}_j \\ + \widetilde{\boldsymbol{\theta}}_W \mathbf{X}_{jW} + \widetilde{\boldsymbol{\theta}}_P \mathbf{X}_{jP} + \mu_{jP} + \widetilde{\varepsilon}_{jP}, \end{aligned} \tag{7}$$

where $\widetilde{\boldsymbol{\theta}}_P = \boldsymbol{\theta}_P + \boldsymbol{\delta}_P$ and $\widetilde{\varepsilon}_{jP} = \widetilde{\varepsilon}_{jW} + \varepsilon_{jP}$. Equation (6) and Equation (7) are explicitly estimated via feasible GLS (FGLS). β_{1P} is identifiable post-estimation by subtracting reduced form coefficients; assuming zero partial correlation between μ_{jP} and female ratio$_j$, it is unbiased.[57]

Both OLS estimation of Equation (3) and FGLS estimation of Equation (6) and Equation (7) require zero partial correlation between μ_{jP} and female ratio$_j$ to obtain a valid β_{1P}.[58] Roughly restated, non-peer review factors must be either independent of its timing (and therefore subsumed in version-invariant fixed effects) or unrelated to gender.[59] Section 3.3.3 evaluates this assumption; briefly, however, I could think of nothing that simultaneously (and convincingly) influences readability, coincides with peer review's timing and correlates with author gender.[60]

3.3.3 Results. Table VII presents results from OLS estimation of Equation (2), FGLS estimation of Equation (6) and Equation (7) and OLS estimation of Equation (3). Since gender bias is possible only when authors' identities are known or can be reasonably guessed, estimates exclude the 279 articles subjected to double-blind review at the *AER* and *QJE* before the internet.[61]

Results in Table VII strongly indicate the readability gap grew precisely while papers were being reviewed. The first column displays β_{1P} from OLS estimation of Equation (2). According to all five scores, women's readability gains outpace men's between versions. Estimates additionally confirm published readability is correlated with draft readability: coefficients on R_{jW} (shown in Appendix D.3) are positive and significant—but only about 0.8.

[57] μ_{jP} may be correlated with $\widetilde{\varepsilon}_{jW}$ via ω_j and/or ε_{jW} without biasing the FGLS estimate of β_{1P} because both are uncorrelated with the explanatory variables in Equation (4) (by assumption) and Equation (6) (by definition).

[58] Unbiased estimation of β_{1P} in Equation (7) requires zero partial correlation between μ_{jP} and female ratio$_j$ after controlling for $\mathbf{X}_{jW}$ and $\mathbf{X}_{jP}$; Equation (3) requires zero partial correlation after controlling for $\mathbf{X}_{jP}$, only.

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

[60] A possible exception is external feedback solicited outside of peer review—*e.g.*, during conferences and seminars. As the next section points out, however, the population of people who provide such feedback overlaps with the population of journal referees. It seems unlikely that this population is biased only in one setting—especially given both settings emphasise gender neutrality.

[61] Excluding these observations does not noticeably impact results or conclusions (for estimates based on the full sample, see Hengel, 2016, p. 18). Two journals—*QJE* and *AER*—employed double-blind review at some point during the time period covered by the data. *QJE* used double-blind procedures until 1 June, 2005. *AER*'s spell began 1 July, 1989 and ended 1 July, 2011. Because a final publication date may substantially lag the actual review date (for an illustration and discussion, see Blank, 1991), I exclude only *AER* articles published after 1992. Economics working papers are generally posted online—and NBER working papers necessarily are—so I assume double-blind review was no longer effective at hiding authors' identities after the internet. Thus, all articles published post-Google's year of incorporation (1998) are included in the sample.

TABLE VII: The impact of gender, specific to peer review

	OLS	FGLS			OLS
	Published article	Working paper	Published article	Difference	Change in score
Flesch Reading Ease	1.33**	2.26**	3.21***	0.95*	0.94
	(0.58)	(1.01)	(1.22)	(0.57)	(0.60)
Flesch-Kincaid	0.52***	0.31	0.76***	0.44**	0.44**
	(0.18)	(0.23)	(0.28)	(0.18)	(0.19)
Gunning Fog	0.52***	0.44*	0.86***	0.42**	0.42**
	(0.19)	(0.24)	(0.29)	(0.19)	(0.20)
SMOG	0.30**	0.33**	0.56***	0.24**	0.24*
	(0.13)	(0.16)	(0.19)	(0.12)	(0.12)
Dale-Chall	0.18***	0.32***	0.45***	0.13**	0.13**
	(0.05)	(0.10)	(0.11)	(0.05)	(0.05)
Editor effects	✓	✓	✓		✓
Journal effects	✓	✓	✓		✓
Year effects	✓	✓	✓		
Journal×Year effects	✓	✓	✓		✓
Quality controls	✓[3]	✓[3]	✓[3]		✓[4]
Native speaker	✓	✓	✓		✓

Notes. Sample 1,801 NBER working papers; 1,799 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles (see Footnote 61). Column one displays coefficients on female ratio (β_{1P}) from estimating Equation (2) directly via OLS (see Appendix D.3 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 (6) and Equation (7), respectively; standard errors clusterd by year and robust to cross-model correlation in parentheses. Their difference (β_{1P}) is shown in column four. Column five displays β_{1P} from OLS estimation of Equation (3); standard errors clustered by year in parentheses. Quality controls denoted by ✓[3] include citation count, max. T_j and max. t_j; ✓[4] includes max. t_j, only (see Footnote 55). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

A less than unit value suggests μ_{jP} exerts downward pressure on R_{jW}'s coefficient, thereby artificially inflating first column figures (see previous section).

Table VII's remaining columns present results from both strategies meant to deal with this bias. Columns 2–4 display FGLS estimates. Coefficients on female ratio$_j$ from Equation (6) ($\widetilde{\beta}_{1W}$) and Equation (7) ($\widetilde{\beta}_{1W} + \beta_{1P}$) are shown in columns two and three, respectively. Female-authored working papers and published articles are both better written—but the readability gap is substantially larger in the latter. Flesch-Kincaid, Gunning Fog and SMOG scores imply immediate peer review accounts for 40–60 percent of the total (biased) effect of female ratio in Equation (7); Flesch Reading Ease and Dale-Chall scores indicate a smaller proportion (30 percent).[62] Column four displays their difference (β_{1P}); it is positive and significant or highly significant for all five scores.

OLS estimates of β_{1P} from Equation (3) are shown in Table VII's final column. Magnitudes are close to FGLS estimates—confirming earlier conclusions—standard errors are slightly higher. Both strategies show a significant increase in the gender readability gap *ex post*. Assuming non-peer review factors are always independent of either its timing or gender, this establishes the desired causal link.[63]

[62]FGLS difference (β_{1P}, column four) divided by the effect in published articles ($\widetilde{\beta}_{1W} + \beta_{1P}$, column three).
[63]The discussion in Footnote 59 also applies to the precise accuracy of the assumption's phrasing used here.

Robustness. Timing independence is the principle independence assumption required to causally link the readability gap with peer review. One external factor in particular may coincide with this timing: feedback women receive in conferences and seminars. Perhaps women tighten prose (before or after submission) in response to audience member remarks? Anecdotal evidence suggests female speakers are given a harder time,[64] although I could find no scientific analysis to support (or contradict) this claim.[65] Nevertheless, most participants are also current (or future) journal referees. Neutral review feedback is inconsistent with non-neutral presentation feedback when originating from the same group.[66]

3.4 Investigating readability over authors' lifetimes

The wider gap post-peer review confirms a causal link with peer review. It does not assure causality with referee scrutiny. In this section, I evaluate the alternatives: women write more clearly because of gender differences in (i) biology/behaviour—*e.g.*, they're more sensitive to referee criticism—or (ii) knowledge about referee expectations—*e.g.*, by overestimating the importance of writing well.

In a dynamic model of authors' decision-making processes, I show that any gap caused exclusively by (i) or (ii) declines with experience. Yet the gap does not decline. It widens. Estimates from pooled subsamples and matching indicate women write more clearly as their publication count increases; men, possibly less so. This pattern of behaviour suggests discrimination—either directly in the form of biased referee scrutiny or indirectly from biased referee assignment (Theorem 1).

3.4.1 Theoretical framework. To organise the analysis, I develop a simple dynamic model of readability's marginal impact on 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 varies on readability.

At stage 0, author i drafts his tth 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 Σ; μ_i is the set of strictly positive probability measures on Σ. Σ and μ_i are known to i.

Let r_{0it} and $\widetilde{r}^s_{0i}$ denote manuscript t's non-negative draft readability and the initial rejection threshold review group $s \in \Sigma$ applies to all papers by author i, respectively. s rejects the paper at stage 0 if

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

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,

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

[64] A related theory is that women receive more critical feedback in conferences and seminars because they present their work more often. In a survey of economists, Sarsons (2016) finds that men and women are equally likely to present co-authored work but women are actually *less* likely to present solo-authored work.

[65] A recent article on Chronicle Vitae discusses the topic and provides specific examples (Baker, 2015). SXSW Interactive (a large technology conference that isn't specifically linked to academia) cancelled two 2015 panel discussions on issues related to gender in response to violent online harassment of the (female) speakers.

[66] Even if this were the case, it implies an entrenched discipline-wide bias.

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

To aid the revision process, s writes a referee report from which i forms expectations about $\widetilde{R}_i^s$ by assigning subjective probabilities $\pi_{1it}^s(R)$ 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 his subsequent expectations about $\widetilde{R}_i^s$ inexact and possibly specious.

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 Σ.[67] Consequently, author i's stage 1 choice of R_{it} maximises his (immediate) subjective expected utility given s,

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

$\Pi_{1it}^s(R_{it})$ is the cumulative sum of $\pi_{1it}^s(R)$ for all $R \leq R_{it}$; u_i is the utility of having a paper accepted in a prestigious journal;[68] $\phi_{i|r_{0it}}(r_{1it}) = \phi_i(R_{it}) - \phi_i(r_{0it})$ and $c_{i|r_{0it}}(r_{1it}) = c_i(R_{it}) - c_i(r_{0it})$ are the satisfaction and cost, respectively, from making changes r_{1it} given the paper's initial readability r_{0it}. ϕ_i is increasing and concave in its arguments, c_i increasing and convex—marginally higher R_{it} generates proportionally less satisfaction but needs more effort when the paper is already well written. $c_i(0)$ and $\phi_i(0)$ are 0.

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

$$\int_\Sigma \Pi_{0it}^s(r_{0it}) v_{1it}^s \, \mathrm{d}\mu_i + \phi_i(r_{0it}) - c_i(r_{0it}), \tag{9}$$

where $\Pi_{0it}^s(r_{0it})$ is the cumulative sum for all $r \leq r_{0it}$ of author i's subjective probabilities $\pi_{0it}^s(r)$ about $\widetilde{r}_{0i}^s$; v_{1it}^s 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 Π_{0it}^s and Π_{1it}^s based on conclusive evidence derived from the choices and outcomes of equivalent peers (Definition 1)[69] and knowledge acquired during his own prior experience in peer review.[70]

Definition 1. *Equivalent authors write identical papers in terms of topic, novelty and quality.*

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

[68] 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.

[69] Specifically, if i observes with probability 1 that in state s an equivalent author k receives an R&R at r_{0k}, then $\Pi_{0it}^s(r) = 1$ for all $r \geq r_{0k}$. Similarly, if i observes with probability 1 that in state s, k is accepted at R_k, then $\Pi_{1it}^s(R) = 1$ for all $R \geq R_k$.

[70] If i is accepted at stage 1 in time t' for review group s, then $\Pi_{1it}^s(R) = 1$ for all $t > t'$ and $R \geq R_{it'}$; otherwise, $\Pi_{1it}^s(R) = 0$ 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 $\Pi_{0it}^s(r) = 1$ for all $t > t'$ and $r \geq r_{0it'}$; otherwise, $\Pi_{it}^s(r) \leq \Pi_{it'}^s(r)$ for all $t > t'$, $r \leq r_{0it'}$ and $s \in \Sigma$.

Equation (8) and Equation (9) incorporate a variety of factors that potentially affect authors' readability choices—editorial standards ($\widetilde{r}_{0it}$ and $\widetilde{R}_{it}$); ambition (u_i); the cost of drafting and revising manuscripts (c_i); an otherwise unexplained intrinsic satisfaction from writing readable papers (ϕ_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, ϕ_i, c_i), (b) editorial standards and/or referee assignment outside his control ($\widetilde{r}_{0it}$, $\widetilde{R}_{it}$, μ_i) or (c) miscellaneous confounding factors mopped by Π^s_{0it} and Π^s_{1it}. Since preferences and costs are time independent, however, an observed increase in i's choice of readability at two separate t distinguishes (a) from the combined impact of (b) and (c).[71] 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 impact on choices declines with experience. Authors may miscalculate referee expectations and misconstrue their reports, but with experience they correct their mistakes. Having ruled out (a) and holding acceptance rates constant, this implies that a persistent readability gap between equivalent peers is caused by (b)—*i.e.*, editorial standards and/or referee assignment beyond authors' control.

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 B.

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.

[71] The analysis in Section 3.3 similarly establishes that (b) and/or (c) are significant factors driving the choice of R_{it}. It cannot, however, distinguish *between* (b) and (c).

Theorem 1's three conditions are sufficient to verify discrimination in academic publishing: 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 referees 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 discrimination. Theorem 1's three conditions confirm the presence of discrimination. They principally rely on two identifying assumptions: (i) i and k are equivalent; (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. By assuming a more specific belief structure at t', Corollary 1 proposes a conservative measure of discrimination's impact on readability choices.

When making revisions, authors choose R_{it} to maximise Equation (8). As shown in Appendix B, $R^\star_i \leq r_{0it}$ where $R^\star_i$ is the R that solves $\phi'_i(R) = c'_i(R)$. Since $R^\star_i$ is i's optimal readability in the absence of peer review and $R^\star_i \leq r_{0it}$, i prefers $R_{it} > r_{0it}$ only if $r_{0it} < \widetilde{R}^s_i + e^s_{1it}$, where e^s_{1it} is his time t error in beliefs about $\widetilde{R}^s_i$. So i revises only when required—and even then, no more than a comfortable minimum to placate referees.

A similar logic governs i's choice of r_{0it}—now picked to maximise Equation (9). i opts for $r_{0it} > R^\star_i$ only if $R^\star_i < \widetilde{r}^s_{0i} + e^s_{0it}$ for at least one s in $\Sigma_{A_{it}}$, where e^s_{0it} is the time t error in i's beliefs about $\widetilde{r}^s_{0i}$. Thus

$$r_{0it} = \max\left\{R^\star_i, \widetilde{r}^{\overline{s}}_{0i} + e^{\overline{s}}_{0it}\right\} \quad \text{and} \quad R_{it} = \max\left\{r_{0it}, \widetilde{R}^s_i + e^s_{1it}\right\}, \tag{10}$$

where $\overline{s}$ is the review group in $\Sigma_{A_{it}}$ for which i believes $\widetilde{r}^s_{0i}$ is highest—*i.e.*, $\overline{s} \in \Sigma_{A_{it}}$ satisfies $\widetilde{r}^s_{0i} + e^s_{0it} \leq \widetilde{r}^{\overline{s}}_{0i} + e^{\overline{s}}_{0it}$ for all $s \in \Sigma_{A_{it}}$.[72]

Define δ^s_{0ik} and δ^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:

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

When $\delta^s_{0ik} \neq 0$ and/or $\delta^s_{1ik} \neq 0$, s employs asymmetric evaluation criteria to i and k's work.[73] Dissimilar authors may call for asymmetric benchmarks—but if i and k are equivalent, they're a form of discrimination. Unfortunately, $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$ are not known to the researcher and R_{it} inconsistently estimates them (Equation (10)). As Corollary 1 shows, however, $R_{it} - R_{kt}$ is *smaller* in magnitude than the true value of stage 1 discrimination by s or stage 0 discrimination by $\overline{s}$.

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

$$R_{it} - R_{kt} \leq D_{ik}, \tag{11}$$

where

$$D_{ik} = \begin{cases} \delta^s_{1ik} & \text{if } r_{0it} < R_{it} \\ \delta^{\overline{s}}_{0ik} & \text{otherwise} \end{cases}.$$

[72] As shown in Theorem 1's proof (Appendix B), i's beliefs about $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$ converge from above. Coupled with Jensen's inequality, this means $\widetilde{r}^s_{0i} + e_{0it}$ and $\widetilde{R}^s_i + e^s_{1it}$ may exceed i's time t expectations of $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$, respectively. At the limit, however, e^s_{0it} and e^s_{1it} converge to 0—so as t increases, this "comfort buffer" declines.

[73] The asymmetry's direction captured in the sign: positive if s is tougher on i; negative otherwise.

Corollary 1 identifies a conservative measure of discrimination's impact on i's readability. It also exposes the toxic denouement of one 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 ($\overline{s}$).

Such is the first externality from even one rotten apple. From i's perspective, $\overline{s}$ spoils the bunch. Bias from $\overline{s}$ destabilises s's attempt to treat i and k fairly. Either i is rejected when assigned to $\overline{s}$ or discrimination by $\overline{s}$ affects i's readability even when i is reviewed by referees who do not discriminate.

Moreover, offsetting unfairness with fairness only works when *everyone* is fair. Asymmetry from one upsets symmetric criteria applied everywhere else, creating endless imbalance when some people just will not be fair. If culture and/or behaviour predicate bias against i and restrain comparable bias against k then, sans intervention, we permanently and unjustly take from i and give to k.[74]

Corollary 1 adds two stronger conditions to Theorem 1. According to the first, i and k must be comparably experienced by time t. Corollary 1 actually applies under the weaker $e^s_{nit} \leq e^s_{nkt}$, $n = 0, 1$ (see its proof in Appendix B), but $R_{it} - R_{kt}$ may overestimate D_{ik} if $e^s_{nkt} < e^s_{nit}$ for all $t > t'$. Nevertheless, $e^s_{nit} - e^s_{nkt}$ converges to 0 as t tends to infinity, so $R_{it} - R_{kt}$ consistently predicts the *direction* of D_{ik} for large enough t.[75]

The second condition precludes s' such that s' is in $\Sigma_{A_{it}}$ but not in $\Sigma_{A_{kt}}$—*e.g.*, because i's utility of acceptance exceeds that of k's. Of course, i's unconditional acceptance rate is not higher than k's (Condition 3), so s' necessarily offsets some other s'' such that—because s'' discriminates against i—s'' is in $\Sigma_{A_{kt}}$ but not in $\Sigma_{A_{it}}$. But $R_{it} - R_{kt}$ may not fully counteract the first effect; Equation (12) does—providing a conservative estimate of D_{ik} under Theorem 1's weaker Condition 3.[76]

$$R_{it} - \max\{R_{it''}, R_{kt}\} \leq D_{ik}. \tag{12}$$

3.4.2 Empirical consistency. If topic, novelty and quality are appropriately controlled for, then discrimination is present when Theorem 1's three conditions hold at large enough t. In this section, I evaluate whether each condition holds, on average, using the entire sample of authors. In Section 3.4.3, I use a matching procedure to identify Theorem 1 and generate a conservative estimate of discrimination's impact on readability (Corollary 1).

Consider first Condition 3—female-authored papers are accepted no more often than male-authored papers. The articles I evaluate have already been accepted, precluding gender analysis of acceptance rates. Section 3.4.3 and Appendix D.8 use lifetime publication counts to partially overcome this. The measure, unfortunately, embodies obvious imperfections.[77]

[74] That is, if cultural and/or behavioural factors mean that $\delta^s_{nik} > 0$ for at least one $s \in \Sigma$, *and* there is no comparable offsetting bias against k *and* education and/or time cannot eliminate δ^s_{nik}, then i is at a permanent disadvantage relative to k.

[75] See also the discussion in Footnote 72 and Section 3.4.3.

[76] Although Equation (12) counteracts the impact of any s' such that s' is in $\Sigma_{A_{it}}$ but not in $\Sigma_{A_{kt}}$, it comes at a cost: Equation (12)'s attenuation bias is much larger than the one generated by Equation (11).

[77] Comparing lifetime publication counts between equivalent authors accounts for most confounding factors except individual productivity—especially factors related to household responsibilities. Greater responsibility at home presumably does not affect readability (other than, perhaps, to push women's scores downward), but it may impact the number of papers women can write. As shown in Section 3.5, however, motherhood responsibilities after childbirth *do not*, in fact, slow women down during the revision process—at least at *Econometrica*.

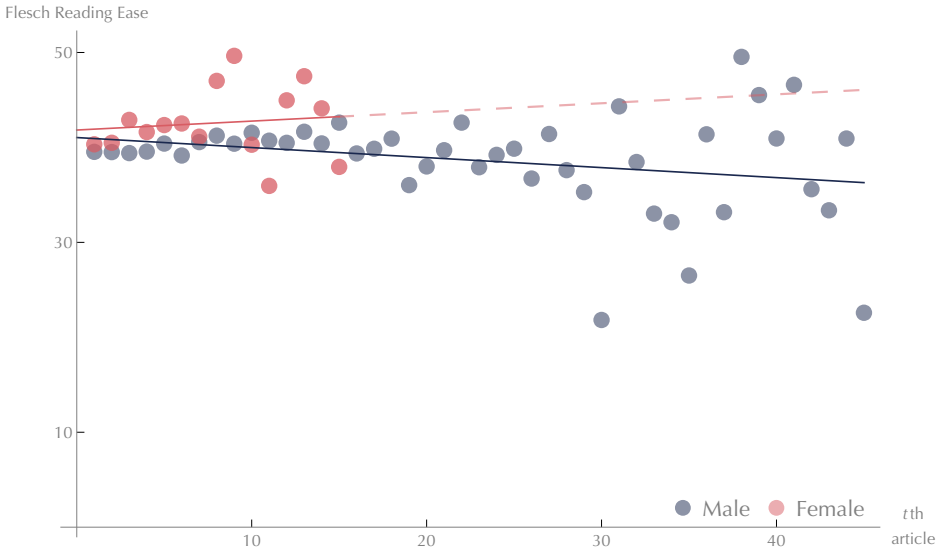


FIGURE III: Readability of authors' tth publication

Notes. Mean Flesch Reading Ease scores grouped by authors' first, second, …, tth, … publication in the data. Lines of best fit are estimated separately for men and women on the grouped averages using OLS. Dotted line indicates out-of-sample forecast (the largest t for a woman is 15; for a man it's 45).

Luckily, gender's impact on acceptance rates has been extensively studied elsewhere. To the best of my knowledge, publication outcomes expose no female advantage anywhere, ever. Blank (1991) found that 12.7 and 10.6 percent of male- and female-authored papers were accepted at the *American Economic Review* , respectively.[78] 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).[79] There are also no gender differences in acceptance rates to NBER's Summer Institute programme (Chari and Goldsmith-Pinkham, 2017).[80] Ceci et al. (2014) provide a much 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).

The data more cleanly identify Conditions 1 and 2. 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 higher (Appendix C). For a man, however, his average and last paper may be more poorly written than the first.

Figure III plots mean Flesch Reading Ease scores grouped by authors' tth article; as the

[78] Women's double-blind acceptance rate was 10 percent (11 percent for men); their single-blind acceptance rate was 11.2 percent (versus 15 percent for men).

[79] The figures presented here aggregate responses in Tables 3 and 4 from Gilbert et al. (1994, p. 141). They average all individual referee recommendations, of which papers usually received several. The authors found no gender difference in final manuscript acceptance rates—although they did find that manuscripts with male corresponding authors were summarily rejected more often (41.7 percent as opposed to 37.4 percent for women).

[80] No gender difference was found in the pooled sample, but male-authored papers submitted to finance workshops were two percent more likely to be accepted; the effect is weakly significant. NBER's annual Summer Institute Programme is a selective three week economics conference.

count increases, men and women diverge.[81] Table VIII tests significance of that divergence by FGLS estimation of Equation (1) (omitting R_{it-1}) on subsamples corresponding to authors' first ($t = 1$), second ($t = 2$), third ($t = 3$), fourth and fifth ($t = 4$–5) and sixth and up ($t \geq 6$) articles published in the journals and time periods covered by the data. Only marginal effects on co-authoring with women for female authors are shown (β_1). Final column is a population-averaged estimate on the pooled sample. Regressions in columns ($t = 1$) to ($t \geq 6$) are weighted by $1/N_j$ (see Section 3.2), standard errors adjusted for two-way clustering on editor and author and corrected for cross-model correlation. Final column estimates are unweighted, error correlations are specified by an auto-regressive process of order one and standard errors are clustered on author.

All figures agree—women write better—but the magnitude and significance of that difference increases as t increases.[82] Between columns ($t = 1$) and ($t = 2$), 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.[83] At higher publication counts, estimates are somewhat smaller than column ($t = 3$)—but still larger than columns ($t = 1$) and ($t = 2$)—although figures are only weakly significant and suffer from very small samples of female authors.[84]

First-time publications are not driving the observed readability gap. Figure III suggests little or no gender difference when $t = 1$; Table VIII backs this up. Coefficients in column ($t = 1$) are imprecise, roughly half the size of those from a pooled regression (last column) and a fraction the size of estimates in columns ($t = 3$), ($t = 4$–5) and ($t \geq 6$). Wald tests (Appendix D.4) reject equality of β_1 in the first and third models 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.

3.4.3 Matching. In light of Theorem 1, the preceding evidence forcefully hints that academic publishing is biased against female economists: on average, female-authored papers are accepted no more often than male-authored papers (Condition 3), yet women improve their writing over time (Condition 2) and write better than men at all t (Condition 1).

Nevertheless, 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 match female to male authors on characteristics that predict the topic, novelty and quality of 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 of higher standards on authors' readability (Corollary 1).

Estimation strategy. 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).

[81] In an earlier version of this paper, I estimated the mean additional contribution each paper makes to an author's readability (Hengel, 2016, pp. 23–24). This analysis included the full set of controls used in Section 3.2. The results and conclusions were similar to those presented here.

[82] See Appendix D.4 for coefficient equality test statistics.

[83] Figures in columns ($t = 2$) and ($t = 3$) of Table VIII are roughly in line with third column estimates in Table VII—on average, $t = 2.7$ for female-authored articles released first as NBER working papers.

[84] Only 40 female authors have 4–5 publications in the data; 28 have six or more. (512 men have 4–5 publications; 545 have more than that.)

TABLE VIII: Gender gap in readability at increasing t

	$t=1$	$t=2$	$t=3$	$t=4$–5	$t\geq 6$	All
Flesch Reading Ease	0.26	1.57*	4.67***	2.73	3.35	1.66**
	(0.63)	(0.91)	(1.14)	(1.97)	(2.14)	(0.72)
Flesch-Kincaid	0.07	0.18	0.82***	0.58	0.62	0.21
	(0.15)	(0.21)	(0.25)	(0.41)	(0.43)	(0.15)
Gunning Fog	0.20	0.38	1.10***	0.83*	0.88	0.44**
	(0.16)	(0.25)	(0.30)	(0.47)	(0.54)	(0.18)
SMOG	0.11	0.27	0.73***	0.64*	0.66*	0.33**
	(0.12)	(0.17)	(0.21)	(0.37)	(0.38)	(0.13)
Dale-Chall	0.07	0.10	0.34***	0.29*	0.43*	0.17**
	(0.06)	(0.08)	(0.12)	(0.17)	(0.24)	(0.07)
No. observations	6,876	2,827	1,674	1,908	2,777	12,013
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓^{5}	✓^{5}	✓^{5}	✓^{5}	✓^{5}	✓^{1}
Native speaker	✓	✓	✓	✓	✓	✓

Notes. β_1 from FGLS estimation of Equation (1) without lagged dependent variable. First column restricts sample to authors' first publication in the data ($t=1$), second column to their second ($t=2$), *etc.* Regressions weighted by $1/N_j$ (see Section 3.2). 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 and max. T_j fixed effects; ✓^{5} includes citation count, only. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

I consider authors "experienced" by $t=3$. Authors with one or two top-four publications are probably 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 $\widetilde{R}_{i3}$, so remaining errors are no longer gender specific: $e^s_{ni3}=e^s_{nk3}$, $n=0,1$ (Corollary 1).[85]

To account for equivalence, I matched every female author with three or more publications (121) to her closest male counterpart (1,553). Matches were made based on the probability of treatment (female) from a probit model with the following co-variates:[86] (1) T_i; (2) mean N_{it}; (3) minimum order in an issue; (4) fraction of papers first-authored by i; (5) maximum citation count; (6) maximum institutional rank; (7) mean publication year; (8) fraction of papers published per decade; (9) fraction of papers published by each journal; and (10) number of articles per primary *JEL* category.[87] Fractions, means, minimums and maximums were calculated over T_i. Co-variate balance pre- and post-match are shown in Appendix D.5. Appendix D.7 lists each matched pair.

$\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$ may be influenced by factors that vary with t: female ratio, journal, year, co-

[85] Recall that $e^s_{nit}-e^s_{nkt}$ converges to 0, so for large enough t Equation (11) and/or Equation (12) predict the direction of D_{ik} even when errors remain gender-specific. (See discussions in the next section.)

[86] The probability of treatment was estimated using the entire sample (771 female authors; 6,105 male authors). Matches were restricted to authors with three or more publications, *ex post*.

[87] I eschewed means in favour of minimum order in an issue, maximum citation count and maximum institutional rank on the assumption that an author's "quality" is principally a function of his best paper.

TABLE IX: $\widehat{R}_{i3} - \widehat{R}_{k3}$ (Condition 1) and $\widehat{R}_{i3} - \widehat{R}_{i1}$ (Condition 2)

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Condition 1					
$\widehat{R}_{i3} - \widehat{R}_{k3}$	8.798***	1.610***	2.629***	1.876***	0.844***
	(1.729)	(0.360)	(0.438)	(0.313)	(0.162)
Condition 2					
$\widehat{R}_{i3} - \widehat{R}_{i1}$ (women)	4.290***	0.892***	1.310***	0.792***	0.004
	(1.544)	(0.316)	(0.388)	(0.285)	(0.133)
$\widehat{R}_{k3} - \widehat{R}_{k1}$ (men)	−3.216**	−0.722**	−1.275***	−0.893***	−0.544***
	(1.397)	(0.304)	(0.349)	(0.243)	(0.115)

Notes. Sample 121 matched pairs (104 and 121 distinct men and women, respectively). $\widehat{R}_{it}$ and $\widehat{R}_{kt}$ are observation-specific readability scores estimated at female ratio equal to 0 for men, 1 for women and $t = 3$ median values of remaining t-dependent co-variates (see Appendix D.5 and text for more details). Figures are weighted by the frequency observations are used in a match. Degrees-of-freedom corrected standard errors in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

author characteristics and stereotypes about authors' institutions. $\widehat{R}_{it}$ accounts for this. It reconstructs R_{it} at female ratio equal to 1 for women, 0 for men and median $t = 3$ values of other co-variates—number of co-authors, institutional rank, institutional rank of the highest ranked co-author, t for the most experienced co-author, publication year, dummies for each journal—using relevant coefficients and residuals from four separate time- and gender-specific regressions on readability.[88] (See Appendix D.6 for regression output.) Throughout the next section (and appropriate appendices), standard errors adjust for the degrees of freedom lost when generating $\widehat{R}_{it}$.[89]

Results. Table IX's first row compares equivalent authors (holding experience constant): senior female economists write more readably than their male counterparts with identical experience. Table IX's last two rows compare authors before and after gaining that experience (holding gender and preferences constant): women write more clearly once they "learn the ropes" in peer review; equivalent men do not. Meanwhile, lifetime publication counts—a crude approximation for acceptance rates—indicate men's more poorly written papers are accepted at least as frequently as women's.[90]

Table IX and average publication counts confirm Section 3.4.2's analysis. Table X goes further. It tests if Conditions 1 and 2 are both satisfied within each matched pair. Its first and second panels display the mean (first column) and standard deviation (second column) of $\underline{D}_{ik}$—Equation (11)'s conservative estimate of D_{ik} (Corollary 1)—and observation counts (third column) 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.[91] Male scores

[88] That is, a dual-authored paper published in 2008 in the *American Economic Review* where $t = 3$ for i and his co-author and their institutions rank 48 and 54, respectively.

[89] Specifically, standard errors are inflated by a factor of 1.2.

[90] See Footnote 77 for a discussion of the limitations of using publication counts to proxy for acceptance rates. Please also refer to Section 3.4.2 for a more thorough review of the (substantial) prior research on gender neutrality and journals' acceptance rates. It too finds no female advantage in journals' acceptance rates.

[91] The co-variates used to generate a match remain relatively balanced when the sample of observations is restricted to $\underline{D}_{ik} \neq 0$ (see Appendix D.5 and the next section for a discussion).

TABLE X: $\underline{D}_{ik}$, 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	18.32	12.94	58	−12.42	10.58	21	6.69***	6.02***
							(1.62)	(1.68)
Flesch Kincaid	3.70	2.68	61	−2.05	2.11	25	1.40***	1.22***
							(0.34)	(0.35)
Gunning Fog	5.11	3.31	62	−3.12	2.57	17	2.23***	2.03***
							(0.42)	(0.44)
SMOG	3.64	2.35	63	−2.44	1.95	16	1.58***	1.44***
							(0.30)	(0.32)
Dale-Chall	1.94	1.30	48	−0.96	0.65	23	0.57***	0.51***
							(0.15)	(0.16)

Notes. Sample 121 matched pairs (104 and 121 distinct men and women, respectively). First and second panels display conditional means, standard deviations and observation counts of $\underline{D}_{ik}$ (Equation (11)) 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.

are subtracted from female scores, so $\underline{D}_{ik}$ is positive in panel one and negative in panel two.

Evidence of discrimination was present in roughly 65 percent of matched pairs—and in three-quarters of those, the member discriminated against was female.[92] Moreover, $\underline{D}_{ik}$ is (on average) almost twice as large (in absolute value) when discrimination is against women.

Figure IV displays $\underline{D}_{ik}$'s distribution across the five scores. Pink bars correspond to matched pairs in which $\underline{D}_{ik}$ is positive (discrimination against women); blue bars reflect those for which $\underline{D}_{ik}$ is negative (discrimination against men).

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 at zero. When women are discriminated against, $\underline{D}_{ik}$ spreads out. Furthermore, instances of obvious discrimination are predominately against women: $\underline{D}_{ik}$ is seven times more likely to be one standard deviation above zero than below it.

Table X's final panel averages $\underline{D}_{ik}$ 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.[93]

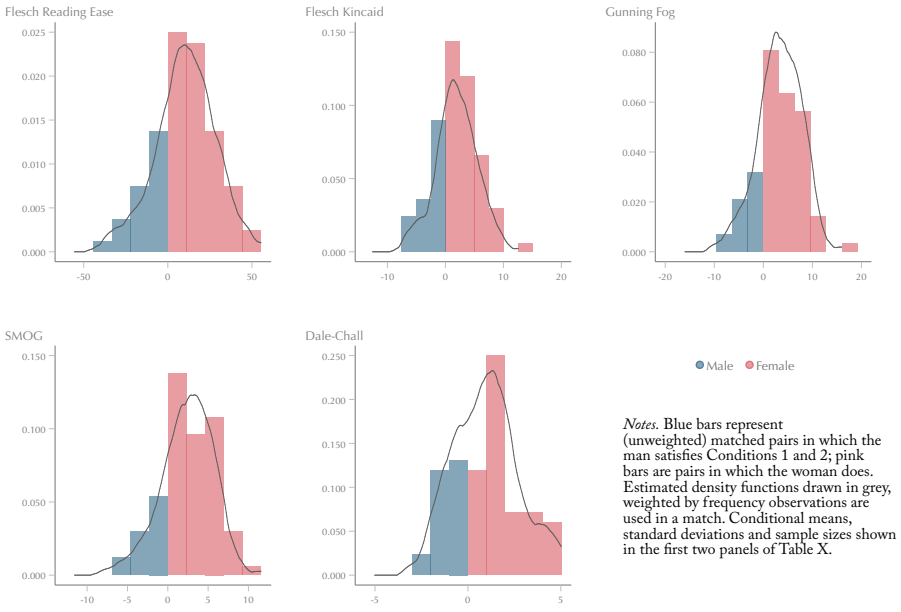
Results confirm conclusions drawn from Figure IV and the first two panels of Table X. Discrimination by editors and/or referees predominately affects female authors. Mean $\underline{D}_{ik}$ is positive and significant in both columns for all five scores. Thanks to higher standards, senior female economists write (at least) nine percent more clearly than they otherwise would.[94]

Appendix D.8 replicates Table X using Equation (12) to estimate D_{ik}. Results are very similar (and conclusions identical) to the analysis presented here.

[92] For 30–40 percent of pairs, neither member satisfied both Conditions 1 and 2, rendering Theorem 1's test for discrimination inconclusive.

[93] 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.

[94] Table X, column (1) divided by the mean male $\widehat{R}_{k3}$ (Appendix D.8).



Notes. Blue bars represent (unweighted) matched pairs in which the man satisfies Conditions 1 and 2; pink bars are pairs in which the woman does. Estimated density functions drawn in grey, weighted by frequency observations are used in a match. Conditional means, standard deviations and sample sizes shown in the first two panels of Table X.

FIGURE IV: Distributions of $\underline{D}_{ik}$, Equation (11)

Robustness. Conclusions drawn from Table X are principally predicated on two assumptions: (i) i and k are equivalent; (ii) t is sufficiently large—*i.e.*, $t > t'$ (e^s_{nit} is on the convergence path to zero for $n = 0, 1$) and any errors in i's beliefs about $\widetilde{r}_{0i}$ and $\widetilde{R}_i$ are sufficiently small.[95] If either is violated, discrimination against women cannot be inferred from an overrepresentation of matched pairs with $\underline{D}_{ik} > 0$.

The first assumption depends on match accuracy. Post-match co-variates are well balanced (Appendix D.5). They remain well balanced—and similar to the matched population—when restricted to pairs satisfying $\underline{D}_{ik} > 0$ and/or $\underline{D}_{ik} < 0$ (Appendix D.5). To facilitate further scrutiny, Appendix D.7 lists the names of economists in each pair.

Matches are sensitive to the choice and construction of variables and the model and method used to estimate propensity scores. Outcomes, however, are not. After controlling for T_i, decade, journal and *JEL* code, matches using alternative variables (*e.g.*, minimum citation counts and mean institutional rank) and specifications (*e.g.*, logit and no replacement) generate similar figures and conclusions.[96]

The second assumption demands a "sufficiently large" t. For diagnosing discrimination, "sufficiently large" means $t' < 3$ and the difference in i and k's error in beliefs at $t = 3$ is smaller than D_{ik}. Forty-eight percent of all women with three or more top publications satisfy Conditions 1 and 2 when compared to equivalent men.[97] Among them, $\underline{D}_{ik}$ is far from zero (Table X, first column): these women write, on average, 29 percent more clearly than

[95] I use "error" and "mistake" to refer to anything that would cause authors to write more (or less) clearly than they would if $\widetilde{r}^s_{0i}$ and $\widetilde{R}^s_i$ were known. This includes actual mistakes in judgement as well as character components—*e.g.*, conscientiousness or risk aversion—that impact beliefs and/or the optimal choice set under uncertainty.

[96] Alternative specifications are not shown, but are available on request (erin.hengel@gmail.com).

[97] Women are the better writers in 73 percent of matched pairs. In 34 percent of those, however, the woman did not improve her writing between $t = 1$ and $t = 3$ (Condition 2), thus rendering Theorem 1's test for discrimination inconclusive.

equivalent men with identical experience.[98] It is unlikely that half of all female economists with three top publications—plus many more second-tier publications and substantial experience refereeing and editing themselves—make mistakes of this magnitude.

Interpreting $\underline{D}_{ik}$ as a causal, conservative estimate of discrimination's impact on readability requires the stronger assumption that $e^s_{ni3} = e^s_{nk3}$.[99] When violated, I can no longer conclude that $\underline{D}_{it}$ conservatively estimates D_{ik}.[100] Nevertheless, $e^s_{nit} - e^s_{nkt}$ is converging to zero and likely very small at $t = 3$. Any upward bias from $e^s_{nkt} < e^s_{nit}$ —*i.e.*, from senior female economists *still* making more mistakes about reviewers' thresholds than equivalent men even after previously publishing two top papers—is probably small and arguably offset by the downward bias already baked into $\underline{D}_{ik}$.[101]

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) $r_{0k3} \leq r_{0i3}$—*i.e.*, i's draft readability is at least as high as k's; and (iii) $r_{0i1} \leq r_{0i3}$—*i.e.*, i's draft readability at $t = 3$ is at least as high as his draft readability at $t = 1$. 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 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 $t = 1$ and $t = 3$ 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 D_{ik}, more generally. First, the previous section reviews the literature on gender neutrality in journals' acceptance rates. Women are not accepted more often than men. In Appendix D.8, I attempt to control for them explicitly by adding the requirement $T_i \leq T_k$ or $T_k \leq T_i$ to categorise matched pairs as discrimination against i or k, respectively. Results are similar; conclusions unchanged. As shown in Section 3.3, women's draft papers are indeed more readable than men's. Appendix A provides further confirmation. It plots the readability of women's and men's draft and published papers over increasing t. Women's drafts are more readable than men's drafts at $t = 3$ *and* more readable than their own earlier drafts at $t = 1$.

3.5 Duration of peer review

"Writing simply and directly only looks easy" (Kimble, 1994, p. 53). An essay's rhetorical competency is highly correlated with the length of time one is given to compose it (Hartvigsen, 1981; Kroll, 1990). Skilled writers spend more time contemplating a writing assignment, brainstorming and editing. They also write fewer words per minute and produce more drafts (Faigley and Witte, 1981; Stallard, 1974).

Since writing simply and directly takes time, one observable repercussion will be prolonged peer review for female authors. To investigate, I turn to *Econometrica*, the only journal

[98] Table X, first column divided by the mean male $\widehat{R}_{k3}$ (Appendix D.8).

[99] $\underline{D}_{ik}$ actually remains a causal, conservative estimate of the impact of discrimination against women under the weaker assumption $e^s_{ni3} \leq e^s_{nk3}$, $n = 0, 1$ (i female, k male). See the proof of Corollary 1 in Appendix B.

[100] Specifically, this assumption is violated if at $t = 3$ the women listed in Appendix D.7 make more (positive) mistakes about $\widetilde{r}^s_{0i}$ and/or $\widetilde{R}^s_i$ than the men they are matched to. For $\underline{D}_{ik}$ to remain a *conservative* estimate of D_{ik}, women's mistakes must be no greater than men's mistakes at $t = 3$.

[101] For a description of this downward bias, see the discussion on Corollary 1 in Section 3.4.1 and the proof of Corollary 1 in Appendix B.

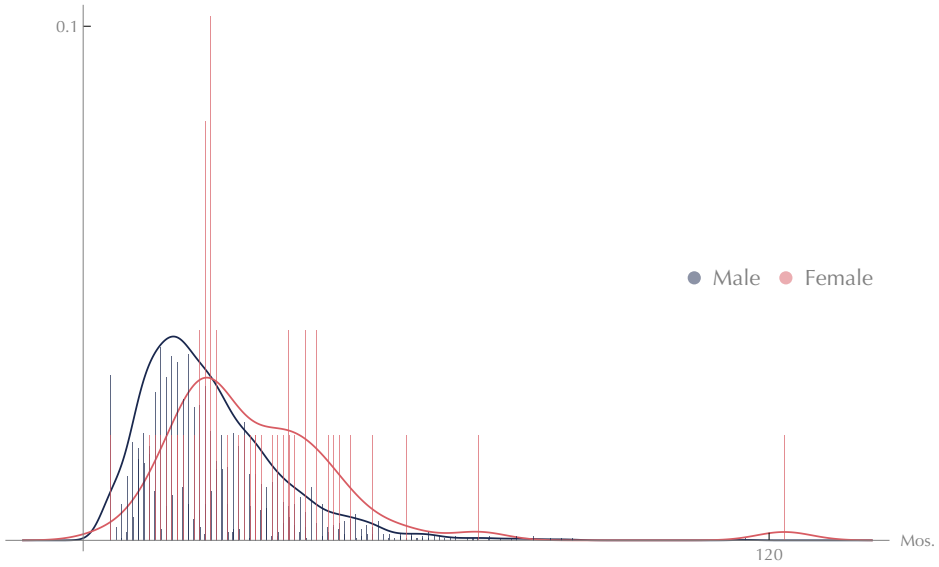


FIGURE V: Distribution of review times at *Econometrica*

Notes. Sample 2,446 articles. Bars are proportional to the number of papers published in *Econometrica* with a given review time (months between first submission and final acceptance). Blue bars represent papers written only by men (2,397); pink bars are papers written only by women (49). Source: *Econometrica*.

to make disaggregated data on the revision process publicly available.

Figure V is a histogram of time (in months) between dates papers are first submitted to and their final revisions received by *Econometrica*'s editorial office. Blue bars represent articles written only by men, pink bars are those just by women. The 180 papers co-authored by men and women are not included.

Since 1950, *Econometrica* published 53 papers authored entirely by women.[102] As Figure V illustrates, their review times disproportionately cluster in the distribution's right tail: articles by women are six times more likely to experience delays above the 75th percentile than they are to enjoy speedy revisions below the 25th.[103]

For a more precise appraisal, I build on a model by Ellison (2002, Table 6, p. 963) and estimate Equation (13):

$$\begin{aligned} \text{revision duration}_j = {} & \beta_0 + \beta_1 \,\text{female ratio}_j + \beta_2 \,\text{mother}_j + \beta_3 \,\text{birth}_j \\ & + \beta_4 \,\max t_j + \beta_5 \,\text{no. pages}_j + \beta_6 \, N_j \\ & + \beta_7 \,\text{order}_j + \beta_8 \,\text{no. citations}_j + \boldsymbol{\theta}\, \mathbf{X}_j + \varepsilon_j, \end{aligned} \tag{13}$$

where mother_j and 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,[104] $\max t_j$ is the number of prior papers published in any of the top four economics journals by article j's most prolific co-author, no. pages_j refers to the page length of the pub-

[102] Submit-accept times were not available for four of these articles (see Section 2).

[103] Despite making up just 2 percent of the sample, one such paper holds the record for longest review: Andrea Wilson's "Bounded Memory and Biases in Information Processing" (November, 2014). Ms. Wilson's paper took a decade to get published.

[104] 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.

TABLE XI: Revision duration at *Econometrica*

	(1)	(2)	(3)	(4)	(5)	(6)
Female ratio	5.290**	6.632***	6.636***	5.541***	6.654***	8.797***
	(2.011)	(2.164)	(2.144)	(2.051)	(2.150)	(2.719)
Max. t_j	−0.163**	−0.169**	−0.164**	−0.164**	−0.163**	−0.169*
	(0.070)	(0.071)	(0.070)	(0.070)	(0.070)	(0.087)
No. pages	0.180***	0.179***	0.178***	0.180***	0.178***	0.206***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.042)
N	1.020**	0.973**	0.963**	1.007**	0.970**	1.149
	(0.443)	(0.442)	(0.443)	(0.443)	(0.443)	(0.698)
Order	0.223**	0.221**	0.218**	0.221**	0.220**	0.496**
	(0.089)	(0.090)	(0.089)	(0.089)	(0.089)	(0.218)
No. citations	0.000	0.000	0.000	0.000	0.000	−0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Mother			−6.660**		−10.934***	−17.672***
			(2.681)		(3.212)	(3.285)
Birth				−2.252	7.579*	12.337**
				(3.360)	(4.167)	(5.588)
Constant	37.708***	37.596***	37.787***	37.692***	37.892***	14.853***
	(2.038)	(2.080)	(2.045)	(2.047)	(2.057)	(2.791)
Editor effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓	✓
JEL (primary) effects						✓
No. observations	2,626	2,610	2,626	2,626	2,626	1,281

Notes. (2) excludes papers authored only by women who gave birth (9 articles) and/or had a child younger than five (16 articles) at some point during peer review. Coefficients from OLS estimation of Equation (13); standard errors clustered by year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

lished article, order_j is the order in which article j appeared in an issue and no. citations_j are the number of subsequent papers citing j.[105]

Table XI displays results across a range of specifications. Column (1) does not control for motherhood or childbirth; (2) drops papers authored 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 childbirth and motherhood; (6) includes fixed effects for primary *JEL* categories.[106]

Every paper published in *Econometrica* undergoes extensive review, but the consistently large and highly significant coefficient on female ratio suggests women bear the worst of it.[107] The average male-authored paper takes 18.5 months to complete all revisions; papers by women need more than half a year longer.[108]

Why? Well, it's not because of motherhood. Yes, giving birth slows down review—responding to referees is apparently put on hold for the first year of a newborn's life—but

[105] I control for all significant factors identified by Ellison (2002). His work evaluates whether author compositional effects contributed to higher mean-accept times at *AER*, *Econometrica*, *JPE*, *QJE* and the *Review of Economic Studies*.

[106] *JEL* classifications are only available for papers published after 1990 (see Section 2); Table XI's column (6) estimates Equation (13) on only half of the data.

[107] This conclusion is robust to altering the age-threshold on mother_j (see Appendix D.9).

[108] Based on results in (5). Male effect estimated with zero female co-authors (standard error 0.102).

having a young child has the opposite effect. A pause for childbirth is expected; a productivity boost from pre-schoolers is not. Perhaps wanting to spend time with the kids motivates women to get organised? Or, maybe the most organised women are the only ones having children? The former suggests motherhood is not the productivity killer it's rumoured to be—at least among highly educated women. The latter implies only superstar women feel academic careers and motherhood are simultaneously manageable.[109] Both interpretations are provocative, but should be made with caution given (i) counter-intuitive results, (ii) obtaining an unbiased estimate of β_2 was *not* this study's objective and (iii) mother$_j$ equals one for only 16 articles in the sample.[110]

As for Table XI's remaining coefficients, all are significant or highly significant and correspond to earlier estimates by Ellison (2002). Longer papers take more time to review, as do papers with more co-authors and those that appear earlier in an issue. Authors with an established publication history and highly cited papers (possibly) enjoy marginally faster reviews.[111]

4 Discussion

A gender readability gap exists. It's still there after including editor, journal and year effects—meaning we cannot blame specific policies or attitudes in the 50s, long since overcome. The gap is unaffected by field controls, so it's not that women research topics that are easier to explain. Perhaps it's caused by factors correlated with gender but actually linked to authors' (or co-authors') competence as economists or fluency in English? If so, institution and native speaker dummies would reduce it. They do not.[112]

The gap grows between first draft and final publication and over the course of women's careers. This precludes systemic bias by article- or author-specific fixed effects—*e.g.*, inborn advantages and one-off improvements in response to external circumstances unrelated to peer review.

It likewise rules out gender differences in (i) biology/behaviour—*e.g.*, sensitivity to referee criticism[113]—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—shouldn't even oversensitive, ill-informed women *eventually* re-examine beliefs… and start acting more like men (Theorem 1)? 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.

[109] A third hypothesis is that referees (possibly responding to editors) demand fewer revisions when women have young children. Because reviewers are unlikely to have this information—based on my own experience, it is remarkably difficult to find out—I (perhaps unfairly) give this interpretation less weight.

[110] The count increases to 17 and 19 articles when mother$_j$'s threshold is defined as children younger than ten and 18, respectively (see Appendix D.9).

[111] Ellison (2002)'s analysis includes a dummy variable for female authorship; it is positive post–1990 but not significant (it is negative and insignificant before that). His paper does not discuss the finding.

[112] 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.

[113] 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).

Women's papers are more likely assigned female referees (Abrevaya and Hamermesh, 2012; Gilbert et al., 1994).[114] If female referees are more demanding critics, clearer writing could reflect their tougher reviews.[115] Women concentrate in particular fields, so it's natural their papers are more often assigned female referees. However, for the readability gap to exist only because of specialisation, controlling for *JEL* classification should explain it.[116] It does not. In fact, even including 718 tertiary *JEL* category dummies has virtually no effect. So if referee assignment is causing the gap, it's only because journals disproportionately refer female-authored papers to the toughest critics. Meaning it isn't referees who are biased—it's editors.[117]

A final alternative is rather uncomfortable. Perhaps female-authored manuscripts deserve more criticism because they aren't as good? 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 institution, total article count, citation counts and published order in an issue—has no effect.[118] 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.

Yet I cannot rule out the possibility that women's work is systematically worse than men's—or that the female and male authors in Section 3.4.3 are not really equivalent. (To decide for yourself, see Appendix D.7.) And if this is true, referees *should* 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.[119]

"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 3.4.1, gender has virtually zero impact on the latter;[120] a review of past

[114]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, although matching does increase in 2000–2008.

[115]It's not so clear whether their reports are any 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).

[116]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.

[117]This is a form of biased referee assignment (Theorem 1). A similar argument contends that female research is more provocative, and more provocative work warrants more scrutiny. If this were true, controlling for *JEL* classification would also reduce (or eliminate) the gap—unless women's work is systematically more provocative even among researchers in very narrow fields. Yet provocative work is (presumably) highly cited work, and there is no discernible gender difference in citation counts (Ceci et al., 2014). Alternatively, perhaps the wider public excessively scrutinises female work, and referees respond similarly to minimise blowback. This explanation assumes a wider public capable of discrediting our work—a view many economists would (privately) disagree with. In any case, economics employs advanced mathematics and technical language, making it especially inaccessible to a layperson.

[118]Published order in an issue refers to the order an article appears in a particular issue (*i.e.*, one for the lead article, two for the second article, *etc.*). This control was introduced as a a set of indicator variables. See Hengel (2016, p. 42 and p. 44) for regression output.

[119]It does seem contradictory, however, that women would be capable of writing better than men—even before referee input (Table VII)—but incapable of producing similar quality research. One is inclined to believe clarity of thought and quality of research to go hand-in-hand, although I am not aware of any study on the topic.

[120]Journals may have a policy of publishing female-authored research over equal (or even better) male work. If

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).

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?[121] Maybe women actually enjoy being poorly informed, overconfident and sensitive to criticism—or (more likely) I may have otherwise misspecified the author's objective function in Section 3.4.1. It is also possible that the statistically significant relationships this paper documents are unfortunate (particularly for me!) flukes.

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, those papers undergo more thorough review. Added scrutiny should improve women's exposition but lengthen review times—as seen in Section 3.5. The rewards from clearer writing are presumably internalised, meaning women gradually improve—which they do, as illustrated in Section 3.4.

Moreover, several studies document a gender difference in critical feedback of similar form—employee performance reviews and student evaluations.[122] 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).[123]

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).[124] 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).

4.1 Open review

Academia's female productivity gap is as stubborn as the business world's pay gap; yet, if every paper a woman writes needs ***six more months*** to finish review, our "Publishing Paradox" seems much less paradoxical.[125]

---

so, acceptance rates are not an unbiased indicator of quality.

[121] 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). Moreover, differences that are present decline—or even reverse—as people age (Weisberg et al., 2011).

[122] No one (to my knowledge) has tested whether men and women receive different critical feedback in peer review reports,

[123] 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).

[124] These conclusions are based on an observational account of the data.

[125] Virtually every study on gender differences in scientific publishing rates find men more productive than women (for a list, see Ceci et al., 2014). It's no different in my data: women published on average 1.7 articles; men managed 2.4—and with far more concentration in the distribution's right tail (for example, 56 men have published 16 or more times in the data, but no woman). Women produce fewer papers even when they don't have any children (Ceci et al., 2014). Appropriate controls for teaching and service do not account for it (Xie and Shauman, 2005), and it isn't a question of time, since female academics work just as many

Is the answer double-blind review? Probably not. Double-blind review cannot stop referees from guessing authors' identities—which they did with surprising accuracy before the internet (Blank, 1991), and presumably perfect accuracy after it.[126]

Instead, eliminate single-blind review, too. A randomised controlled trial at the *British Journal of Psychiatry* suggests referee reports are better quality and less abusive when identities are known (Walsh et al., 2000). Posting them online—as the *British Medical Journal* does—virtually guarantees continuous, independent audits by outside researchers.[127] Worries that reviews are less critical and/or relationships are strained are either unfounded or alleviated by the deep pool of referees common to general interest journals (van Rooyen et al., 2010; van Rooyen et al., 1999). Open review does incur costs—some people refuse to participate and those that don't spend marginally more time drafting reports (van Rooyen et al., 1999; Walsh et al., 2000)[128]—but if more accountability promotes fairer outcomes, ethical arguments in its favour should outweigh minor practical concerns.

5 Conclusion

This paper makes a curious discovery: female-authored articles in top economics journals are better written. After examining the difference, I conclude that higher standards applied by editors and/or referees are primarily to blame.

No prior study has uncovered convincing evidence of gender bias in journal acceptance rates. It's encouraging that sex is irrelevant to publication outcomes, but that does not mean it has no effect on the process—or on the productivity of female academics. When female authors endure unfair criticism in referee reports, clearer writing and longer review times follow. With less time to spend on new projects, research output slows down.

Higher standards impose a quantity vs. quality tradeoff that not only reconciles academia's "Publishing Paradox", but also rationalises many instances of female output. Work that is evaluated more critically at any point in the production process will be systematically better (holding prices fixed) or systematically cheaper (holding quality fixed). This reduces women's wages—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—and distorts measurements of female productivity—billable hours and client revenue decline; female lawyers appear less productive than they truly are.

My findings also emphasise the importance of transparency and monitoring. Unlike referee reports, journal acceptance rates are easy to measure and frequently audited; both

hours as men (Ceci et al., 2014; Ecklund and Lincoln, 2011).

[126] In an earlier version of this paper, I show that the gender readability gap is actually *higher* when papers are evaluated blindly (for results and discussion, see Hengel, 2015, pp. 64–67).

[127] The *BMJ* posts reviewers' signed reports, authors' responses and the original manuscript on its website. No documentation is posted for rejected papers, but doing so may be beneficial: (i) A very public review implies a very public rejection; concern for one's reputation could reduce the number of low quality submissions. (ii) The onus of discovering mistakes would be shared with the wider economics community. (iii) Other journals can make publication decisions based on posted reviews—possibly reducing time spent refereeing for the discipline, as a whole. Women may receive greater scrutiny online—as they do at the *Guardian* (Gardiner et al., 2016)—but the difference can be mitigated if comments are non-anonymous, made only by verified members of an appropriate professional society and continuously (and publicly) audited for bias in quantity and quality of feedback.

[128] Each study employed a different research design; nevertheless, both estimate roughly 12 percent of reviewers decline to participate because they oppose open peer review while signing reports increases time spent on the review by 25 minutes. When referees were told their signed reviews might be posted online, time rose by an additional half hour and refusal rates were much higher (55 percent) (van Rooyen et al., 2010).

factors foster accountability, which encourages neutrality (Foschi, 1996). Monitoring referee reports is difficult but not impossible—especially if peer review were open. Several science and medical journals not only reveal referees' identities, they also post reports online. Quality does not decline (it may actually increase), referees still referee (even those who initially refuse) and the extra 25–50 minutes spent reviewing seems tolerable (van Rooyen et al., 2010; van Rooyen et al., 1999; Walsh et al., 2000).

Finally, the topic of my study is narrow, but its methodology has wider applications. To the best of my knowledge, this paper is the first (in economics) to identify discrimination using the choices and behaviours of those discriminated against. Although applied to a specific context—peer review—the identifying logic equally suits any situation where people repeatedly receive and act on biased feedback. Moreover, this study is also the first to uncover subtle group differences with readability scores.[129] These scores are not new—all are extensively tested with well-documented properties—but their use is mostly confined to determining whether text is appropriate for intended audiences.[130] As this paper demonstrates, however, readability scores are also effective tools to evaluate asymmetry anywhere ideas are communicated orally or in writing and large amounts of source material are easily obtainable: journalism, speeches, student essays, business plans, Kickstarter campaigns, *etc.* Research potential is substantial.

[129] Ali et al. (2010) identified readability scores as useful tools for social scientists. In a large scale analysis of news content, they found stories on sports (male dominated) and entertainment (female dominated) most readable. Stempel (1981) reports similar findings in popular U.S. newspapers.

[130] Long and Christensen (2011), Lehavy et al. (2011) and Thörnqvist (2015) use readability scores in interesting, non-conventional ways. The former investigates whether a legal brief's Flesch Reading Ease score is correlated with its success on appeal (it is not); the latter two use readability measures to proxy for complex information in financial reports, finding less readable material is less informative (Lehavy et al., 2011), especially for non-sophisticated investors (Thörnqvist, 2015). Since releasing the first version of this working paper (September, 2015) research using readability scores has ballooned. See Benoit et al. (2017) for a review of more recent research.

Appendices

A Indirect effect of higher standards

Section 3.4.1 implies a tradeoff: address referees' demands during peer review—but risk desk rejection—or anticipate them before peer review—and risk wasting time. As a final exercise, I investigate gender differences in how this tradeoff is made.

Figure A.1 compares papers pre- and post-review at increasing publication counts. Solid circles denote NBER draft readability; arrow tips reflect readability in the final, published versions of those same papers; dashed lines trace readability as papers undergo peer review. Standard errors from various differences are shown in Table A.1. Figures are based on FGLS estimation of Equation (A.14) (see Section 3.3):

$$R_{j_{it}m} = \beta_0 + \beta_1 \text{ female ratio}_j + \beta_2 \text{ female ratio}_j \times t_i + \beta_3 t_i + \boldsymbol{\theta}\, \mathbf{X}_j + \varepsilon_j, \qquad \text{(A.14)}$$

where $m = W, P$ for working papers and published articles, respectively, and $\mathbf{X}_j$ is a vector of observable controls: editor, journal, year, journal and year interactions, English fluency dummies and quality controls—citation count and max. T_j. Since t_i is author-specific, I disaggregate the data by duplicating each article N_j times; to account for duplicate articles, regressions are weighted by $1/N_j$ (see Section 3.2).[131]

Figure A.1 suggests female economists initially underestimate referees' expectations: without experience, their writing improves during peer review;[132] with experience, they write more clearly before peer review. Women's draft readability increases between $t = 1$ and $t = 2$—and then again between $t = 2$ and $t = 3$. Consequently, women make fewer changes during peer review in $t = 2$ than in $t = 1$; changes shrink further by $t = 3$.

Women's pattern of behaviour both resembles and differs from men's. Draft and final readability scores for male-authored papers remain relatively constant over increasing t. Unlike women's, men's approach does not radically change with experience: they consistently overestimate referee demands pre-peer review to minimise changes made in peer review.[133]

This strategy mirrors women's at later t. Economists who anticipate demands are desk rejected less often; economists who don't enjoy more free time, all things equal.[134] Figure A.1 implies little—if any—gender difference in this tradeoff. Decisions by junior economists may reflect inexperience, but decisions by senior economists should not. Senior economists are familiar with peer review; their choices express optimal tradeoffs with full information (for discussion and justification, see Section 3.4.1). Figure A.1 suggests both men and women sacrifice time to increase acceptance rates.

[131] Results and conclusions based on unweighted regressions—or by replacing t_i with max. t_j and *not* duplicating articles—are very similar or identical to those presented here. Regression output from alternative specifications available on request (erin.hengel@gmail.com).

[132] Assuming no gender difference in acceptance rates at $t = 3$ and given evidence that women are held to higher standards documented earlier, Figure A.1 suggests—but does not prove—that manuscripts by junior female economists are disproportionately rejected.

[133] Consistent with Table VI, readability may actually decline during peer review. As discussed in Section 3.3, this may be an artifact specific to abstracts, which are edited for length in addition to readability. Alternatively, writing (too) well upfront satisfies the review group with the highest initial readability threshold. Because referee reports reveal s (and therefore $\widetilde{R}_i^s$), a readability decline after receiving an R&R indicates that a majority of groups have laxer standards. This explanation is consistent with the theoretical model in Section 3.4.1.

[134] Alternatively, if desk rejection rates are gender neutral, authors subject to higher standards will undergo more arduous peer review. Greater scrutiny would therefore replace higher desk rejection rates when editors (or even referees) monitor and implement a policy of gender neutral acceptance rates.

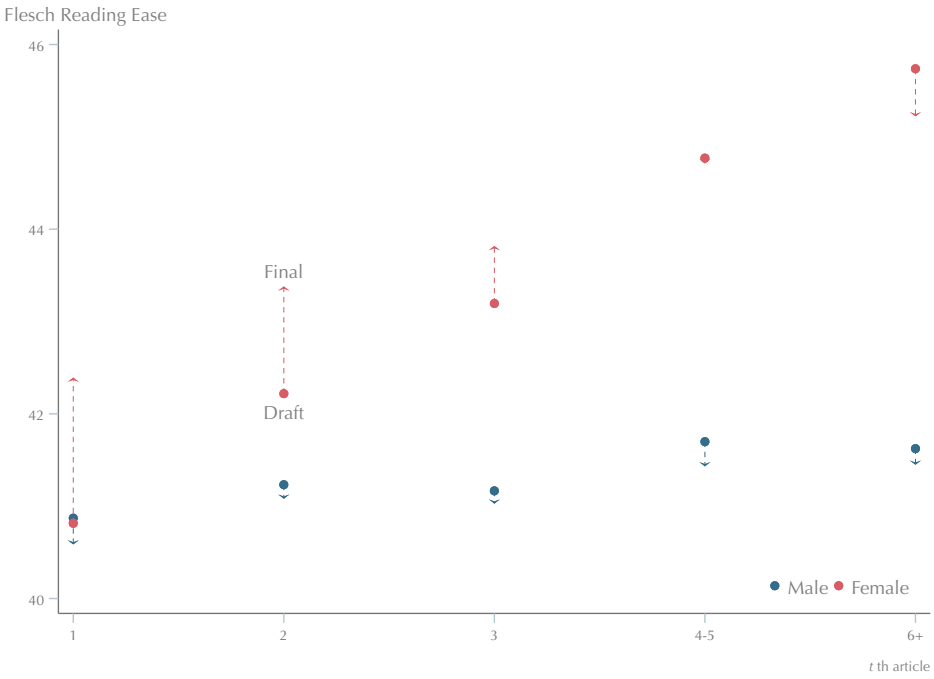


FIGURE A.1: Readability of authors' tth publication (draft and final versions)

Notes. Sample 4,289 observations; 1,988 and 1,986 distinct NBER working papers and published articles, respectively; 1,840 distinct authors. Flesch Reading Ease marginal mean scores for authors' first, second, third, 4th–5th and sixth and up publications in the data. Solid circles denote estimated readability of NBER working papers from FGLS estimation of Equation (A.14); arrow tips show the estimated readability in published versions of the same papers. Controls are: editor, year, journal, journal and year interactions, English fluency dummies and quality controls (citation count and max. T_j). Regression weighted by $1/N_j$. Pink represents women co-authoring only with other women; blue are men co-authoring only with other men.

The first panel of Table A.1 displays the magnitude and standard errors of the contemporaneous marginal effect of peer review ($R_{jP} - R_{jW}$) for men and women over increasing t. Figures correspond to the length of the dotted lines in Figure A.1. The difference in that length, represented in the third row, approximates the relative tradeoffs men and women make. For publications $t = 1$ and $t = 2$, differences are large and significant; in publications three and up, however, they're fairly small. Indeed, by publications $t = 4$–5 and $t = 6+$, men and women mostly choose to address referee concerns prior to peer review, corroborating analysis based on Figure A.1.

Assuming—as other evidence suggests—that men's and women's papers are accepted at identical rates, this unfortunately means senior female economists work harder before submission to achieve the same outcome post submission.[135] The gender gap in readability scores at $t = 6+$ is over four points on the Flesch Reading Ease scale (Table A.1). Senior female economists write approximately ten percent more clearly than men—a figure which roughly corresponds to the causal effect of discrimination estimated in Section 3.4.3.

Moreover, male and female economists write equally well at the start of their careers. There is no gender difference in draft readability at $t = 1$. This suggests—although it cannot conclusively prove—that male and female economists begin their academic careers with the

[135]The smaller gap at $t = 1$ may correspond to higher rejection rates for papers authored by junior female economists. See also Footnote 134 for an alternative interpretation in which acceptance rates are identical but scrutiny is not.

TABLE A.1: Readability of authors' tth publication (draft and final versions)

	$t = 1$	$t = 2$	$t = 3$	$t = 4$–5	$t \geq 6$
Predicted $R_{jP} - R_{jW}$					
Women	1.58**	1.16**	0.62	−0.06	−0.52
	(0.62)	(0.57)	(0.65)	(0.81)	(1.05)
Men	−0.28	−0.15	−0.14	−0.27*	−0.17
	(0.17)	(0.10)	(0.08)	(0.14)	(0.19)
Difference	1.86***	1.31**	0.76	0.21	−0.34
	(0.68)	(0.63)	(0.72)	(0.91)	(1.14)
Marginal effect of female ratio					
Draft paper	−0.06	0.99	2.03**	3.07***	4.11***
	(1.24)	(0.95)	(0.82)	(0.89)	(1.14)
Published article	1.81	2.30***	2.79***	3.28***	3.77***
	(1.19)	(0.86)	(0.77)	(0.99)	(1.37)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Quality controls	✓[3]	✓[3]	✓[3]	✓[3]	✓[3]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 4,289 observations; 1,988 and 1,986 distinct NBER working papers and published articles, respectively; 1,840 distinct authors. Panel one displays magnitude of predicted $R_{jP} - R_{jW}$ (the contemporaneous effect of peer review) for women and men over increasing publication count (t). Panel two estimates the marginal effect of an article's female ratio ($\beta_1 + \beta_2$), separately for draft papers and published articles. Figures from FGLS estimate of Equation (A.14). Quality controls denoted by ✓[3] include citation count, max. T_j and max. t_j. Standard errors clustered by editor and robust to cross-model correlation in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

same information about the peer review process. Unfortunately, that information is accurate only for men.

The second panel of Table A.1 displays the marginal effect of female ratio in draft and published papers. Draft readability increasingly contributes to the published article gap: less than zero percent at $t = 1$, 43 percent at $t = 2$, 73 percent at $t = 3$, 94 percent at $t = 4$–5 and over 100 percent for $t = 6+$. The total gap, however, remains far more stable—between 2–3 Flesch Reading Ease points. This suggests that female economists are held to relatively constant—albeit higher—standards throughout their careers.

Figure A.1 supports Theorem 1's implicit assumption that female authors learn about referees' thresholds over time. If the payoff from lucid exposition is high, people will catch on—either by internalising explicit comments on text readability in referee reports from earlier papers or making the (un)conscious connection that acceptance rates are higher—or review times are faster—when text is clearer. Applying that payoff only to women yields a succinct explanation for the gap's observed growth.

Although Table A.1 concurs, the first panel viewed in isolation gives a different, more orthodox impression. It suggests that the readability gap declines over increasing t. This narrow view favours alternative explanations—*e.g.*, sensitivity, poor information and/or justified statistical discrimination—over bias by referees and/or editors. Only when complemented by Figure A.1 do we fully appreciate that the smaller gap *in* peer review is completely offset by a wider gap *before* peer review.

This raises a broader concern: responses to discrimination may superficially conceal it. Female economists adjust to biased treatment in ways that partially—or even totally—confuse underlying discrimination with voluntary choice. Studies that do not account for this may underestimate the phenomenon, find it does not exist or even conclude bias against men.

For similar reasons, common performance controls may discount discrimination in equations that relate wages (and other labour market outcomes) to gender. Controlling for performance is undeniably import; yet just as important is our judgement and measurement of that performance. Higher standards in narrow dimensions that fail to contribute to the observed value of output will lower superficial measures of female productivity and confound gender differences in labour market outcomes.

B Proofs

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. Throughout, $\{(r_{0it}, R_{it})\}$ represents the sequence of readability choices made by author i for all t. $R_i^\star$ is defined as the R that solves $\phi_i'(R) = c_i'(R)$. 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

$$\int_\Sigma \left(\pi_{0it}^s(r_{0it}) v_{1it}^s + \Pi_{0it}^s(r_{0it}) \frac{\partial v_{1it}^s}{\partial r_{0it}} \right) \mathrm{d}\mu_i + \phi_i'(r_{0it}) - c_i'(r_{0it}) = 0, \tag{B.15}$$

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,

$$\begin{aligned} \frac{\partial v_{1it}^s}{\partial r_{0it}} &= \pi_{1it}^s(R_{it}) u_i + \phi_i'(R_{it}) - c_i'(R_{it}) - \phi_i'(r_{0it}) + c_i'(r_{0it}) \\ &= \frac{\partial v_{1it}^s}{\partial r_{1it}} + c_i'(r_{0it}) - \phi_i'(r_{0it}). \end{aligned} \tag{B.16}$$

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 (B.15) evaluated at $r_{0it} = R_i^\star$ is correspondingly equivalent to

$$\int_\Sigma \left(\pi_{0it}^s(r_{0it}) v_{1it}^s + \Pi_{0it}^s(r_{0it}) \frac{\partial v_{1it}^s}{\partial r_{1it}} \right) \mathrm{d}\mu_i. \tag{B.17}$$

v_{1it}^s is non-negative;[136] 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 (B.17) is non-negative. Since $c_i'(r) < \phi_i'(r)$ for all $r < R_i^\star$, the left-hand side of Equation (B.15) 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,

$$\lim_{r_0 \to \infty} \Pi_{0it}^s(r_0) v_{1it}^s = \overline{\Pi}_{0it}^s \overline{\Pi}_{1it}^s u_i, \tag{B.18}$$

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 $r > R_i^\star$,

$$\lim_{r_0 \to \infty} \left\{ \int_\Sigma \Pi_{0it}^s(r_0) v_{1it}^s \, \mathrm{d}\mu_i + \phi_i(r_0) - c_i(r_0) \right\} = -\infty. \tag{B.19}$$

[136] 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,

$$\sup \left\{ \underset{r_{0it}}{\operatorname{argmax}} \int_\Sigma \Pi^s_{0it}(r_{0it}) v^s_{1it} \, \mathrm{d}\mu_i + \phi_i(r_{0it}) - c_i(r_{0it}) \right\} < \infty, \tag{B.20}$$

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 $\Pi^s_{1it}(R_{it}) \leq 1$

$$\begin{aligned} &\Pi^s_{1it}(R_{it}) u_i + \phi_i(R_{it}) - \phi_i(r_{0it}) - c_i(R_{it}) + c_i(r_{0it}) \\ &\quad \leq u_i + \phi_i(R_{it}) - c_i(R_{it}) \end{aligned} \tag{B.21}$$

Equation (B.22) 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 (B.22) 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 Σ 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

$$\pi^s_{1it}(R^s_{it}) = \frac{1}{u_i} \left(c'_i(R^s_{it}) - \phi'_i(R^s_{it}) \right). \tag{B.22}$$

Let π^s_{1i} denote the terminal value of π^s_{1it} as t tends to ∞. π^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 violation of Assumption X.

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 (B.22) implies

$$\begin{aligned}
&\lim_{t\to\infty}\left|\pi^s_{1it+1}(R^s_{it+1})-\pi^s_{it}(R^s_{it})\right| \\
&\quad=\lim_{\varepsilon\to 0}\frac{1}{u_i}\Big(\left[c_i'(R^s_{it}+\varepsilon)-c_i'(R^s_{it})\right]-\left[\phi_i'(R^s_{it}+\varepsilon)-\phi_i'(R^s_{it})\right]\Big) \\
&\quad=\frac{1}{u_i}\left(c_i''(R^s_i)-\phi_i''(R^s_i)\right), \qquad \text{(B.23)}
\end{aligned}$$

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 (B.23) is strictly positive. According to Equation (B.23), $\{\pi^s_{1it}\}$ does not converge, a contradiction.

Consider now the proposition that $r_{0i}\le r_{0it}$ 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 $r^s_{1it+1}\le r^s_{1it}$ for all $s\in\Sigma$.

At time t, the author chooses r_{0it}. This choice is governed by the first-order condition in Equation (B.15):

$$K+\mu^s_i\left(\pi^s_{0it}(r_{0it})v^s_{1it}+\Pi^s_{0it}(r_{0it})\frac{\partial v^s_{1it}}{\partial r_{0it}}\right)=c_i'(r_{0it})-\phi_i'(r_{0it}) \qquad \text{(B.24)}$$

where μ^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 (B.16), Equation (B.24) is equivalent to

$$K+\mu^s_i\pi^s_{0it}(r_{0it})v^s_{1it}=\Big(1-\mu^s_i\Pi^s_{0it}(r_{0it})\Big)\Big(c_i'(r_{0it})-\phi_i'(r_{0it})\Big). \qquad \text{(B.25)}$$

If $r^s_{1it}=0$ then $r^s_{1it+1}=0$, and $\partial v^s_{1it}/\partial r_{1it}=\pi^s_{1it}(R^s_{it})u_i$.[137] In this case, Equation (B.24) is equivalent to

$$K+\mu^s_i\Big(\pi^s_{0it}(r_{0it})v^s_{1it}+\Pi^s_{0it}(r_{0it})\pi^s_{1it}(R^s_{it})u_i\Big)=c_i'(r_{0it})-\phi_i'(r_{0it}). \qquad \text{(B.26)}$$

By the monotone convergence theorem, $\{v^s_{1it}\}$ and $\{\Pi^s_{0it}\}$ converge.[138] 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 $\widetilde{r}^s_{0i}\le r_{0it}$; if $r_{0i}\le\widetilde{r}^s_{0i}$, then

$$\lim_{t\to\infty}\Pi^s_{0it}(r_{0it})=\lim_{t\to\infty}\sum_{r\in\Omega_t}\pi^s_{0it}(r)=\pi^s_{0i}(r_{0i}), \qquad \text{(B.27)}$$

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

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

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

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

and

$$\begin{aligned}
&\lim_{t\to\infty} \left|\mu_i^s u_i \left(\Pi_{0it+1}^s(r_{0it+1})\pi_{1it+1}^s(R_{it+1}^s) - \Pi_{0it}^s(r_{0it})\pi_{1it}^s(R_{it}^s)\right)\right| \\
&\quad = \mu_i^s\, u_i \left(\lim_{t\to\infty} \Pi_{0it+1}^s(r_{0it+1}) \lim_{t\to\infty} \pi_{1it+1}^s(R_{it+1}^s) - \lim_{t\to\infty} \Pi_{0it}^s(r_{0it}) \lim_{t\to\infty} \pi_{1it}^s(R_{it}^s)\right) \\
&\quad\quad = 0.
\end{aligned}$$

For the moment, assume there exists t'' such that for all $r \in (r_{0it''}, r_{0i})$, K is constant.[139] Thus, changes over time to the left-hand sides of Equation (B.25) and Equation (B.26) converge to 0. Yet the right-hand sides of Equation (B.25) and Equation (B.26) 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 Π_{0i}^s is the finite limit of $\{\Pi_{0it}^s\}$, while

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

is strictly greater than 0, where convergence of $\{r_{0it}\}$ guarantees that for all (sufficiently small) $\varepsilon > 0$ there exists $r_{0it+1} = r_{0it} + \varepsilon$.[140] 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 (B.25) and Equation (B.26) 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. □

[139] 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.

[140] 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 (B.27) shows is not possible.

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: $(r'_{0i}, R_i^{s\prime})$ and $(r''_{0i}, R_i^{s\prime\prime})$. Denote their respective convergent subsequences by $\{(r'_{0it}, R_{it}^{s\prime})\}$ and $\{(r''_{0it}, R_{it}^{s\prime\prime})\}$. Given the concavity of ϕ_i and convexity of c_i, a unique readability at each stage maximises Equation (8) and Equation (9) for fixed Π_{0it}^s and Π_{1it}^s. Thus, $r'_{0i0} = r''_{0i0}$ and $R_{i0}^{s\prime} = R_{i0}^{s\prime\prime}$ at time 0.

Assume at time t the author has chosen $r'_{0il} = r''_{0il}$ and $R_{il}^{s\prime} = R_{il}^{s\prime\prime}$ for all $l < t$; thus, $\Pi_{0it}^{s\prime}(r) = \Pi_{0it}^{s\prime\prime}(r)$ and $\Pi_{1it}^{s\prime}(R) = \Pi_{1it}^{s\prime\prime}(R)$ for all r and R, so the author chooses $r'_{0it} = r''_{0it}$ and $R_{it}^{s\prime} = R_{it}^{s\prime\prime}$ at time t as well. By the axiom of induction, $\{(r'_{0it}, R_{it}^{s\prime})\} = \{(r''_{0it}, R_{it}^{s\prime\prime})\}$ for all t so (r_{0i}, R_i^s) is unique.[141] Since the choice of s was arbitrary 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*

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 $\widetilde{r}_{0i}^s = \widetilde{r}_{0k}^s$, $\widetilde{R}_i^s = \widetilde{R}_k^s$ *and* $\mu_i^s = \mu_k^s$ *for all* $s \in \Sigma$*, then*

$$\int_\Sigma \mathbf{1}_{0k}^s(r_{0kt})\mathbf{1}_{1k}^s(R_{kt})\,\mathrm{d}\mu_k < \int_\Sigma \mathbf{1}_{0i}^s(r_{0it})\mathbf{1}_{1it}^s(R_{it})\,\mathrm{d}\mu_i. \tag{B.28}$$

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 $\Sigma_{A_{it}}$, $R_{kt}^q < R_{it}^q$. As a result,

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

and

$$\mathbf{1}_{0k}^s(r_{0kt})\mathbf{1}_{1k}^s(R_{kt}^s) \leq \mathbf{1}_{0i}^s(r_{0it})\mathbf{1}_{1i}^s(R_{it}^s) = 1 \text{ for } s = q. \tag{B.29}$$

If I show that the inequality in Equation (B.29) is strict, then Equation (B.31) is true. By way of a contradiction, assume it holds as an equality. Thus, $\widetilde{R}_i^q \leq R_k^q < R_i^q$, where $R_{kt}^q \to R_k^q$ and $R_{it}^q \to R_i^q$ (Lemma 4). Together with $R_i^\star \leq r_{0it''} < R_i^q$, this implies

$$\lim_{\varepsilon \to 0-} \Pi_{1i}^q(R_i^q + \varepsilon) < 1.^{142} \tag{B.30}$$

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

$$\lim_{N_k \to \infty} \frac{N_{A_k}}{N_k} = \mu_i^q, \tag{B.31}$$

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

[142] That is, $\Pi_{0i}^q(R) = 1$ for all $R \geq R_i^q$. Because he chose $R_i^\star \leq R_{it''} < R_i^q$ 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 $R_i^q > \max\{R_{it''}, R_k^q\}$, it must be because there is no smaller R that satisfies Equation (B.22). This is only possible if there is a jump discontinuity in Π_{0i}^q at R_i^q, as illustrated in Equation (B.30).

where N_{A_k} and N_k are author k's accepted and total paper counts, respectively. Because i updates Π^s_{1it} when he observes with probability 1 that in state s, k is accepted at some $R \neq R^s_i$ (see Footnote 69), Equation (B.31) 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 (B.29) 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. I first show that Equation (11) conservatively estimates D_{ik} when $\Sigma_{A_{it}} \subset \Sigma_{A_{kt}}$. Let $r_{0it} < R_{it}$. From Equation (10) and the definition of δ^s_{1ik},

$$\begin{aligned} R_{it} - R_{kt} &= \widetilde{R}^s_i + e_{1it} - \max\left\{R^\star_k, \widetilde{r}^{\overline{s}_k}_{0k} + e_{0kt}, \widetilde{R}^s_k + e_{1kt}\right\} \\ &\leq \widetilde{R}^s_i - \widetilde{R}^s_k + e_{1it} - e_{1kt} \\ &= \delta^s_{1ik} + e_{1it} - e_{1kt}. \end{aligned} \tag{B.32}$$

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 (10) and the definition of δ^s_{0ik} instead imply:

$$\begin{aligned} R_{it} - R_{kt} &= \max\left\{R^\star_i, \widetilde{r}^{\overline{s}_i}_{0i} + e_{0it}\right\} - \max\left\{R^\star_k, \widetilde{r}^{\overline{s}_k}_{0k} + e_{0kt}, \widetilde{R}^s_k + e_{1kt}\right\} \\ &\leq \max\left\{R^\star_i, \widetilde{r}^{\overline{s}_i}_{0i} + e_{0it}\right\} - \widetilde{r}^{\overline{s}_k}_{0k} - e_{0kt}, \end{aligned} \tag{B.33}$$

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 (B.33) is equivalent to

$$\begin{aligned} R_{it} - R_{kt} &\leq \widetilde{r}^{\overline{s}_i}_{0i} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt} \\ &= \delta^{\overline{s}_i}_{0ik} + \widetilde{r}^{\overline{s}_i}_{0k} - \widetilde{r}^{\overline{s}_k}_{0k} + e_{0it} - e_{0kt}. \end{aligned} \tag{B.34}$$

$e_{0it} = e_{0kt}$ 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 (B.34) implies $R_{it} - R_{kt} \leq \delta^{\overline{s}_i}_{0ik}$ if $R_{it} = r_{0it}$. Meanwhile, Equation (B.32) implies $R_{it} - R_{kt} \leq \delta^s_{1ik}$ if $r_{0it} < R_{it}$.

It remains to show that Equation (12) 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 δ^s_{1ik}. $R_{it} - R_{it''}$ is a conservative estimate of R_{ik}. Thus, all is proved.

□

C *Average first, mean and final paper scores*

Table C.2 displays authors' average readability scores for their first, mean and final papers. Grade-level scores (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (see Section 2.1). Sample excludes authors with fewer than three publications.

TABLE C.2: Average first, mean and final paper scores

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Average first paper score					
Women	39.20	−13.81	−17.36	−15.18	−11.00
	(1.152)	(0.238)	(0.289)	(0.210)	(0.097)
Men	39.37	−13.77	−17.54	−15.35	−11.00
	(0.307)	(0.072)	(0.082)	(0.055)	(0.026)
Average mean score					
Women	41.20	−13.36	−16.92	−14.92	−10.91
	(0.720)	(0.147)	(0.185)	(0.135)	(0.067)
Men	39.59	−13.69	−17.42	−15.27	−11.02
	(0.186)	(0.043)	(0.048)	(0.033)	(0.016)
Average final paper score					
Women	41.99	−13.10	−16.58	−14.66	−10.90
	(1.060)	(0.215)	(0.253)	(0.182)	(0.107)
Men	39.53	−13.71	−17.41	−15.24	−11.08
	(0.325)	(0.080)	(0.090)	(0.059)	(0.026)

Notes. Sample 1,674 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 Section 2.1). Standard errors in parentheses.

D Supplemental output

D.1 Table IV, journal and male effects. Table D.3 shows male effects from the regressions described and presented in Table IV. Effects estimated at a female ratio of zero and observed values for other co-variates. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (Section 2.1). Table D.3 shows the coefficients on the journal dummies in column (2), Table IV. They compare *AER*'s readability to the readability of *Econometrica*, *JPE* and *QJE*.

TABLE D.3: Table IV, male effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Flesch Reading Ease	39.59	39.59	39.60	39.60	39.58	40.13	40.29
	(0.037)	(0.037)	(0.038)	(0.037)	(0.038)	(0.058)	(0.086)
Flesch-Kincaid	−13.72	−13.72	−13.72	−13.72	−13.73	−13.48	−13.46
	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.013)	(0.017)
Gunning Fog	−17.46	−17.46	−17.46	−17.46	−17.47	−17.16	−17.12
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.015)	(0.020)
SMOG	−15.28	−15.28	−15.28	−15.28	−15.28	−15.10	−15.07
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.011)	(0.015)
Dale-Chall	−11.00	−11.00	−11.00	−11.00	−11.00	−11.03	−11.03
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.008)
Editor effects	✓	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓	✓
Year effects		✓	✓	✓	✓	✓	✓
Journal×Year effects			✓	✓	✓	✓	✓
Institution effects				✓	✓	✓	✓
Quality controls					✓[1]	✓[1]	✓[1]
Native speaker					✓	✓	✓
JEL (primary) effects						✓	
JEL (tertiary) effects							✓

Notes. 9,122 articles in (1)–(5); 5,216 articles in (6); 5,777 articles—including 561 from *AER Papers & Proceedings* (see Footnote 46)—in (7). Figures correspond to the male effects from regression results presented in Table IV. Effects estimated at a female ratio of zero and observed values for other co-variates. Quality controls denoted by ✓[1] include citation count and max. T_j fixed effects. Standard errors clustered on editor in parentheses.

TABLE D.4: Journal readability, comparisons to *AER*

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Econometrica	−12.48***	−4.44***	−4.26***	−2.63***	−0.66***
	(1.93)	(0.41)	(0.47)	(0.38)	(0.16)
JPE	−5.69***	−4.01***	−3.42***	−1.84***	0.18
	(1.93)	(0.41)	(0.47)	(0.38)	(0.16)
QJE	1.47**	−0.04	0.28***	0.19***	0.27***
	(0.63)	(0.14)	(0.09)	(0.07)	(0.05)

Notes. Figures are the estimated coefficients on the journal dummy variables from (2) in Table IV. 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 Table V, male effects. Table D.5 displays total male effects—*i.e.*, the total effect for men co-authoring only with other men—from the regressions presented in Table V. Effects estimated at a female ratio of zero and observed values for other co-variates. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (see Section 2.1).

TABLE D.5: Table V, male effects

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
Male effect	39.79	−13.63	−17.37	−15.23	−11.01
	(0.146)	(0.032)	(0.037)	(0.026)	(0.012)
N_j	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Journal×Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓
Quality controls	✓[1]	✓[1]	✓[1]	✓[1]	✓[1]
Native speaker	✓	✓	✓	✓	✓

Notes. Sample 9,186 observations (2,827 authors). Figures correspond to the male effects from regression results presented in Table V (first-differenced, IV estimation of Equation (1), Arellano and Bover (1995) and Blundell and Bond (1998)). Effects estimated at a female ratio of zero and observed values for other co-variates. Quality controls denoted by ✓[1] include citation count and max. T_j fixed effects. Regressions weighted by $1/N_j$; standard errors adjusted for two-way clustering on editor and author (in parentheses).

D.3 Table VII (first column), full output. Table D.6 estimates Equation (2) via OLS. The first row displays coefficients on the working paper score, R_{jW}. The second row is the coefficient on female ratio (β_{1P}), also shown in the first column of Table VII. Remaining rows present estimated coefficients from the other (non-fixed effects) control variables: Max. t_j and Max. T_j—contemporaneous and lifetime publication counts for article j's most prolific co-author, respectively—number of citations and a dummy variable equal to one if article j is authored by at least one native (or almost native) English speaker.

TABLE D.6: Table VII (first column), full output

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
R_{jW}	0.833***	0.755***	0.773***	0.790***	0.841***
	(0.022)	(0.038)	(0.036)	(0.028)	(0.016)
Female ratio	1.329**	0.518***	0.517***	0.304**	0.179***
	(0.580)	(0.177)	(0.188)	(0.129)	(0.054)
Max. t_j	0.008	0.005	0.008	0.005	−0.004
	(0.072)	(0.018)	(0.019)	(0.012)	(0.004)
Max. T_j	0.016	0.002	0.001	0.000	0.003
	(0.054)	(0.012)	(0.013)	(0.008)	(0.003)
No. citations	−0.001**	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Native speaker	−0.306	0.009	0.023	0.012	−0.057*
	(0.378)	(0.149)	(0.176)	(0.104)	(0.031)
Constant	16.969***	−2.179***	−2.307***	−2.170***	−0.613***
	(0.971)	(0.609)	(0.683)	(0.476)	(0.212)
Editor effects	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Year×Journal effects	✓	✓	✓	✓	✓

Notes. Sample 1,801 NBER working papers; 1,799 published articles. Estimates exclude 279 pre-internet double-blind reviewed articles (see Footnote 61). Coefficients from OLS regression of Equation (2). First row is the coefficient on R_{jW}; second row is β_{1P}, and corresponds to results presented in the first column of Table VII. Coefficients on quality controls (citation counts, max. T_j and max. t_j) also shown. Standard errors clustered on editor (in parentheses). ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

D.4 Table VIII, equality test statistics and male effects. Table D.7 displays χ^2 test statistics from Wald tests of β_1 (Equation (1)) equality across estimation results in Table VIII. Table D.8 shows male effects from the regressions described and presented in Table VIII. Effects estimated at a female ratio of zero and observed values for other co-variates. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (Section 2.1).

TABLE D.7: Table VIII, equality test statistics

	$t = 1$ vs. 2	$t = 1$ vs. 3	$t = 1$ vs. 4–5	$t = 1$ vs. ≥ 6	$t = 2$ vs. 3
Flesch Reading Ease	1.708	12.213	1.505	2.028	5.000
Flesch-Kincaid	0.237	8.470	1.254	1.282	5.080
Gunning Fog	0.537	7.860	1.348	1.397	4.385
SMOG	0.872	7.782	1.664	2.060	3.777
Dale-Chall	0.089	4.169	1.688	2.013	2.669

Notes. χ^2 test statistics from Wald tests of β_1 (Equation (1)) equality across estimation results in Table VIII.

TABLE D.8: Table VIII, male effects

	$t = 1$	$t = 2$	$t = 3$	$t = 4$–5	$t \geq 6$	All
Flesch Reading Ease	39.58	39.55	39.63	39.57	39.82	39.71
	(0.060)	(0.106)	(0.146)	(0.153)	(0.201)	(0.149)
Flesch-Kincaid	−13.71	−13.69	−13.65	−13.63	−13.58	−13.65
	(0.012)	(0.022)	(0.030)	(0.033)	(0.040)	(0.035)
Gunning Fog	−17.43	−17.40	−17.36	−17.35	−17.28	−17.38
	(0.015)	(0.027)	(0.036)	(0.040)	(0.051)	(0.038)
SMOG	−15.25	−15.24	−15.23	−15.23	−15.18	−15.24
	(0.011)	(0.020)	(0.026)	(0.029)	(0.038)	(0.026)
Dale-Chall	−11.02	−11.03	−11.03	−11.05	−11.03	−11.01
	(0.006)	(0.008)	(0.012)	(0.012)	(0.016)	(0.012)
No. observations	6,876	2,827	1,674	1,908	2,777	12,013
N_j	✓	✓	✓	✓	✓	✓
Editor effects	✓	✓	✓	✓	✓	✓
Journal effects	✓	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓	✓
Journal×Year effects						✓
Institution effects	✓	✓	✓	✓	✓	✓
Quality controls	✓[5]	✓[5]	✓[5]	✓[5]	✓[5]	✓[1]
Native speaker	✓	✓	✓	✓	✓	✓

Notes. Figures correspond to the male effects from regression results presented in Table VIII (FGLS estimation of Equation (1) without lagged dependent variable). First column restricts sample to authors' first publication in the data ($t = 1$), second column to their second ($t = 2$), *etc.* Regressions weighted by $1/N_j$ (see Section 3.2). 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 and max. T_j fixed effects; ✓[5] includes citation count, only. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

D.5 Section 3.4.3, co-variate balance. Table D.9 compares co-variate balance pre- and post-match. 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 1,553 male authors with three or more publications. The first column of the second panel ("Post-match means") displays (weighted) averages for the 104 male authors matched with a female author. Table D.10, Table D.11 and Table D.12 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 between $\underline{D}_{ik} > 0$ (discrimination against women) and $\underline{D}_{ik} < 0$ (discrimination against men) samples; both resemble averages in the matched sample.

TABLE D.9: Pre- and post-matching summary statistics

		Pre-match means			Post-match means		
	Women	Men	Difference	t	Men	Difference	t
T	4.55	5.90	−1.35	−3.47	4.65	−0.11	−0.32
Avg. N_{it}	2.20	2.09	0.11	1.99	2.29	−0.09	−1.06
Min. order in issue	2.69	2.43	0.26	1.50	2.88	−0.18	−0.67
% first authored by i	1.83	3.34	−1.51	−1.58	3.44	−1.61	−1.43
Max. citations	267.07	406.62	−139.56	−1.78	278.75	−11.68	−0.20
Max. inst. rank	49.26	44.42	4.83	2.72	47.71	1.54	0.74
Avg. year	2003.48	1995.45	8.03	6.89	2002.15	1.33	1.17
Fraction of articles per decade							
1950–59	0.00	0.01	−0.01	−1.57	0.00	0.00	
1960–69	0.00	0.04	−0.04	−2.87	0.01	−0.01	−0.88
1970–79	0.01	0.11	−0.09	−4.72	0.02	0.00	−0.27
1980–89	0.08	0.18	−0.10	−4.36	0.12	−0.04	−1.43
1990–99	0.19	0.21	−0.02	−0.99	0.17	0.02	0.52
2000–09	0.41	0.26	0.15	5.90	0.39	0.02	0.51
2010–15	0.31	0.20	0.11	4.19	0.30	0.01	0.32
Fraction of articles per journal							
AER	0.39	0.25	0.14	5.52	0.36	0.02	0.57
Econometrica	0.17	0.34	−0.17	−5.12	0.17	0.00	−0.03
JPE	0.18	0.24	−0.07	−2.62	0.19	−0.01	−0.36
QJE	0.27	0.17	0.10	4.78	0.28	−0.01	−0.25
Fraction of articles per *JEL* code							
A General	0.04	0.02	0.02	1.58	0.02	0.02	0.95
B Methodology	0.00	0.02	−0.02	−1.44	0.02	−0.02	−1.53
C Quant. methods	0.64	0.81	−0.17	−1.03	0.52	0.13	0.78
D Microeconomics	1.64	1.79	−0.15	−0.69	1.54	0.10	0.43
E Macroeconomics	0.58	0.62	−0.05	−0.37	0.39	0.18	1.42
F International	0.39	0.31	0.08	0.81	0.28	0.11	0.90
G Finance	0.60	0.52	0.07	0.68	0.42	0.17	1.14
H Public	0.45	0.36	0.10	1.09	0.59	−0.13	−1.01
I Health, welfare, edu	0.88	0.34	0.53	5.35	0.90	−0.03	−0.13
J Labour	1.26	0.76	0.49	3.39	1.47	−0.21	−0.89
K Law and econ	0.20	0.14	0.06	1.14	0.29	−0.09	−1.12
L Industrial org	0.73	0.57	0.16	1.46	0.63	0.09	0.63
M Marketing/accounting	0.17	0.13	0.04	0.92	0.19	−0.03	−0.33
N Economic history	0.29	0.14	0.15	2.73	0.26	0.03	0.28
O Development	0.86	0.52	0.34	2.58	0.93	−0.07	−0.34
P Economic systems	0.08	0.09	−0.01	−0.22	0.09	0.00	−0.08
Q Agri., environment	0.18	0.12	0.06	1.20	0.22	−0.04	−0.50
R Regional, transport	0.17	0.16	0.01	0.16	0.21	−0.04	−0.70
Z Special topics	0.16	0.10	0.06	1.50	0.29	−0.13	−1.74

Notes. Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics (121 female authors, 1,553 male authors). Second panel shows post-match summary statistics (104 male authors). t-values for differences reported in columns four and seven.

TABLE D.10: 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	4.92	4.61	0.32	0.66	4.85	4.71	0.14	0.30
Avg. N_{it}	2.24	2.29	−0.05	−0.48	2.21	2.28	−0.07	−0.78
Min. order in issue	2.70	2.61	0.09	0.26	2.94	2.38	0.56	1.78
% first authored by i	3.63	2.10	1.53	1.09	2.34	3.80	−1.46	−1.03
Max. citations	294.41	309.42	−15.01	−0.20	215.62	342.51	−126.90	−1.90
Max. inst. rank	47.96	47.43	0.53	0.20	47.52	49.86	−2.34	−1.00
Avg. year	2002.03	2003.75	−1.72	−1.29	2001.99	2003.75	−1.77	−1.35
Fraction of articles per decade								
1950–59	0.00	0.00	0.00		0.00	0.00	0.00	
1960–69	0.00	0.01	−0.01	−1.00	0.00	0.01	−0.01	−0.90
1970–79	0.01	0.01	0.00	−0.08	0.01	0.02	−0.01	−0.55
1980–89	0.14	0.07	0.07	1.77	0.12	0.05	0.07	2.07
1990–99	0.18	0.18	0.00	−0.10	0.19	0.17	0.02	0.54
2000–09	0.41	0.38	0.03	0.67	0.42	0.41	0.01	0.14
2010–15	0.26	0.35	−0.09	−1.78	0.26	0.33	−0.08	−1.69
Fraction of articles per journal								
AER	0.35	0.42	−0.07	−1.56	0.36	0.41	−0.05	−1.07
Econometrica	0.14	0.15	−0.01	−0.15	0.18	0.16	0.01	0.33
JPE	0.22	0.16	0.05	1.35	0.19	0.16	0.03	0.82
QJE	0.29	0.27	0.02	0.55	0.27	0.26	0.00	0.05
Fraction of articles per *JEL* code								
A General	0.03	0.04	−0.01	−0.45	0.03	0.02	0.01	0.45
B Methodology	0.03	0.00	0.03	1.42	0.01	0.00	0.01	1.00
C Quant. methods	0.51	0.48	0.03	0.15	0.62	0.51	0.10	0.64
D Microeconomics	1.61	1.70	−0.09	−0.32	1.63	1.69	−0.06	−0.23
E Macroeconomics	0.48	0.51	−0.03	−0.16	0.45	0.49	−0.03	−0.24
F International	0.35	0.35	0.00	0.00	0.26	0.30	−0.05	−0.36
G Finance	0.53	0.54	−0.01	−0.06	0.37	0.67	−0.30	−1.67
H Public	0.63	0.51	0.13	0.74	0.57	0.55	0.02	0.14
I Health, welfare, edu	1.09	0.92	0.16	0.54	1.15	1.02	0.13	0.41
J Labour	1.57	1.33	0.24	0.78	1.53	1.56	−0.02	−0.08
K Law and econ	0.28	0.30	−0.03	−0.24	0.21	0.42	−0.21	−2.01
L Industrial org	0.67	0.80	−0.13	−0.69	0.63	0.76	−0.13	−0.76
M Marketing/accounting	0.24	0.22	0.03	0.23	0.24	0.17	0.07	0.71
N Economic history	0.25	0.34	−0.09	−0.69	0.24	0.31	−0.07	−0.58
O Development	1.08	0.73	0.34	1.21	1.14	0.80	0.34	1.14
P Economic systems	0.10	0.11	−0.01	−0.18	0.14	0.09	0.05	0.63
Q Agri., environment	0.20	0.18	0.03	0.30	0.33	0.13	0.20	2.17
R Regional, transport	0.19	0.20	−0.01	−0.19	0.15	0.26	−0.10	−1.65
Z Special topics	0.27	0.20	0.06	0.70	0.22	0.26	−0.03	−0.38

Notes. Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics. t-values for differences reported in columns four and seven.

TABLE D.11: 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	4.96	4.42	0.54	1.10	5.09	4.46	0.63	1.24
Avg. N_{it}	2.28	2.24	0.03	0.35	2.28	2.26	0.02	0.25
Min. order in issue	2.94	2.56	0.38	1.12	2.89	2.58	0.30	0.89
% first authored by i	3.14	3.04	0.10	0.06	3.23	3.46	−0.24	−0.15
Max. citations	234.84	340.37	−105.53	−1.46	237.73	340.77	−103.04	−1.42
Max. inst. rank	47.76	48.33	−0.57	−0.22	47.38	48.63	−1.25	−0.48
Avg. year	2001.99	2003.64	−1.65	−1.25	2002.00	2003.76	−1.76	−1.35
Fraction of articles per decade								
1950–59	0.00	0.00	0.00		0.00	0.00	0.00	
1960–69	0.00	0.01	0.00	−0.13	0.00	0.01	0.00	−0.13
1970–79	0.01	0.01	−0.01	−0.55	0.01	0.02	−0.01	−0.88
1980–89	0.13	0.07	0.06	1.53	0.13	0.06	0.06	1.78
1990–99	0.19	0.18	0.01	0.26	0.19	0.17	0.02	0.53
2000–09	0.42	0.41	0.01	0.22	0.42	0.42	0.00	−0.07
2010–15	0.25	0.32	−0.07	−1.43	0.25	0.33	−0.07	−1.43
Fraction of articles per journal								
AER	0.35	0.42	−0.07	−1.43	0.35	0.41	−0.06	−1.35
Econometrica	0.16	0.17	−0.01	−0.15	0.16	0.16	0.00	0.04
JPE	0.20	0.17	0.03	0.82	0.20	0.18	0.02	0.51
QJE	0.29	0.25	0.04	0.95	0.29	0.25	0.04	0.92
Fraction of articles per *JEL* code								
A General	0.03	0.03	0.00	0.00	0.03	0.03	0.00	0.00
B Methodology	0.01	0.00	0.01	1.00	0.01	0.00	0.01	1.00
C Quant. methods	0.52	0.66	−0.14	−0.76	0.54	0.63	−0.09	−0.48
D Microeconomics	1.58	1.58	0.00	0.00	1.59	1.57	0.03	0.09
E Macroeconomics	0.47	0.47	0.00	0.00	0.42	0.54	−0.13	−0.84
F International	0.24	0.25	−0.01	−0.11	0.24	0.33	−0.09	−0.64
G Finance	0.47	0.56	−0.09	−0.47	0.52	0.53	−0.01	−0.07
H Public	0.57	0.46	0.11	0.93	0.70	0.46	0.24	1.41
I Health, welfare, edu	1.30	0.90	0.41	1.21	1.33	0.89	0.44	1.32
J Labour	1.61	1.32	0.29	1.01	1.71	1.29	0.42	1.34
K Law and econ	0.25	0.32	−0.06	−0.60	0.25	0.32	−0.06	−0.60
L Industrial org	0.61	0.68	−0.08	−0.45	0.65	0.73	−0.09	−0.51
M Marketing/accounting	0.29	0.11	0.18	1.71	0.29	0.10	0.19	1.84
N Economic history	0.22	0.35	−0.14	−1.10	0.20	0.35	−0.15	−1.20
O Development	1.11	0.82	0.29	0.92	1.09	0.81	0.28	0.88
P Economic systems	0.14	0.10	0.04	0.48	0.14	0.10	0.04	0.48
Q Agri., environment	0.28	0.15	0.13	1.35	0.27	0.18	0.09	0.92
R Regional, transport	0.16	0.24	−0.08	−1.14	0.19	0.24	−0.05	−0.74
Z Special topics	0.28	0.22	0.06	0.67	0.29	0.22	0.08	0.80

Notes. Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics. t-values for differences reported in columns four and seven.

TABLE D.12: Co-variate post-match balance when $\underline{D}_{ik} \neq 0$

	Dale-Chall			
	Discrimination			
	Against women	Against men	Difference	t
T	5.04	4.90	0.14	0.25
Avg. N_{it}	2.29	2.26	0.03	0.29
Min. order in issue	2.77	2.54	0.24	0.71
% first authored by i	1.80	3.26	−1.46	−1.08
Max. citations	205.15	395.77	−190.62	−2.36
Max. inst. rank	46.48	50.23	−3.75	−1.45
Avg. year	2001.59	2002.99	−1.40	−0.95
Fraction of articles per decade				
1950–59	0.00	0.00	0.00	
1960–69	0.01	0.00	0.01	1.42
1970–79	0.02	0.01	0.02	1.33
1980–89	0.13	0.10	0.03	0.83
1990–99	0.17	0.19	−0.02	−0.52
2000–09	0.37	0.40	−0.03	−0.59
2010–15	0.28	0.29	−0.01	−0.19
Fraction of articles per journal				
AER	0.37	0.39	−0.01	−0.30
Econometrica	0.15	0.12	0.02	0.51
JPE	0.23	0.20	0.03	0.65
QJE	0.26	0.30	−0.04	−0.82
Fraction of articles per *JEL* code				
A General	0.00	0.01	−0.01	−1.00
B Methodology	0.01	0.01	0.00	0.00
C Quant. methods	0.45	0.54	−0.08	−0.59
D Microeconomics	1.58	1.79	−0.21	−0.77
E Macroeconomics	0.58	0.52	0.06	0.31
F International	0.44	0.23	0.21	1.37
G Finance	0.41	0.59	−0.18	−1.07
H Public	0.66	0.45	0.21	1.16
I Health, welfare, edu	1.21	1.04	0.17	0.47
J Labour	1.55	1.69	−0.14	−0.39
K Law and econ	0.21	0.34	−0.13	−1.18
L Industrial org	0.58	0.70	−0.13	−0.72
M Marketing/accounting	0.24	0.18	0.06	0.48
N Economic history	0.23	0.32	−0.10	−0.70
O Development	1.30	0.93	0.37	0.99
P Economic systems	0.20	0.08	0.11	1.24
Q Agri., environment	0.28	0.18	0.10	0.90
R Regional, transport	0.21	0.20	0.01	0.20
Z Special topics	0.32	0.23	0.10	0.96

Notes. Sample restricted to authors with three or more publications. First panel shows pre-match summary statistics. t-values for differences reported in columns four and seven.

D.6 Section 3.4.3, $\widehat{R}_{it}$ regression output. Table D.9 displays output from time- and gender-specific regressions used to generate $\widehat{R}_{it}$. (Output for male authors at $t = 1$ not shown.)

TABLE D.13: Regression output generating $\widehat{R}_{it}$

	Flesch Reading Ease			Flesch-Kincaid		
	Women		Men	Women		Men
	$t = 1$	$t = 3$	$t = 3$	$t = 1$	$t = 3$	$t = 3$
Female ratio	4.36	3.89	0.01	−0.09	0.74	1.40
	(7.68)	(5.96)	(7.56)	(1.59)	(1.21)	(1.65)
N	1.33	0.14	−0.32	−0.01	−0.07	−0.03
	(2.47)	(1.89)	(1.34)	(0.51)	(0.38)	(0.29)
Inst. rank	−0.03	−0.15*	0.03	0.00	−0.03	−0.02
	(0.11)	(0.08)	(0.06)	(0.02)	(0.02)	(0.01)
Max. inst. rank	0.07	0.18	−0.09	0.01	0.03	−0.02
	(0.13)	(0.11)	(0.07)	(0.03)	(0.02)	(0.02)
Max. t_j	−0.57*	−0.23	−0.06	−0.06	0.00	−0.06
	(0.34)	(0.29)	(0.29)	(0.07)	(0.06)	(0.06)
Year	0.21	0.03	0.20	0.04	0.03	0.04
	(0.13)	(0.13)	(0.13)	(0.03)	(0.03)	(0.03)
Econometrica	−3.03	−4.25	−0.06	−1.37*	−0.80	−0.73
	(3.90)	(3.10)	(3.43)	(0.81)	(0.63)	(0.75)
JPE	1.05	1.63	5.66*	0.19	0.16	0.18
	(3.39)	(3.34)	(3.21)	(0.70)	(0.68)	(0.70)
QJE	4.95	0.38	5.50**	0.38	−0.31	0.64
	(3.05)	(2.66)	(2.62)	(0.63)	(0.54)	(0.57)
Constant	−389.75	−23.30	−364.60	−97.19*	−75.18	−83.72
	(265.17)	(262.33)	(250.38)	(55.07)	(53.33)	(54.61)

Notes. Sample 121 female authors; 104 male authors. Sample restricted to matched authors. See Section 3.4.3 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.

TABLE D.14: Regression output generating $\widehat{R}_{it}$

	Gunning Fog			SMOG			Dale-Chall		
	Women		Men	Women		Men	Women		Men
	$t=1$	$t=3$	$t=3$	$t=1$	$t=3$	$t=3$	$t=1$	$t=3$	$t=3$
Female ratio	−0.85	1.14	3.07	−0.36	0.90	2.43*	0.28	0.31	0.82
	(1.93)	(1.50)	(1.91)	(1.40)	(1.11)	(1.35)	(0.61)	(0.58)	(0.68)
N	−0.10	−0.05	0.00	−0.01	0.01	0.00	0.08	−0.05	0.07
	(0.62)	(0.47)	(0.34)	(0.45)	(0.35)	(0.24)	(0.19)	(0.18)	(0.12)
Inst. rank	0.01	−0.02	−0.03**	0.00	−0.02	−0.02**	−0.01	−0.02**	−0.02***
	(0.03)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Max. inst. rank	0.01	0.02	−0.02	0.01	0.02	−0.02	0.01	0.02	0.01
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Max. t_j	−0.11	0.00	−0.01	−0.10	−0.01	0.01	−0.04	−0.02	0.01
	(0.08)	(0.07)	(0.07)	(0.06)	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)
Year	0.04	0.02	0.05	0.04	0.01	0.04	0.02**	−0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Econometrica	−1.65*	−1.32*	−0.68	−0.78	−0.96*	−0.40	−0.61*	−0.45	0.22
	(0.98)	(0.78)	(0.87)	(0.71)	(0.58)	(0.61)	(0.31)	(0.30)	(0.31)
JPE	−0.11	0.29	0.66	−0.04	0.40	0.62	−0.11	0.25	0.57*
	(0.85)	(0.84)	(0.81)	(0.62)	(0.62)	(0.57)	(0.27)	(0.33)	(0.29)
QJE	0.65	−0.42	1.15*	0.61	−0.23	0.81*	0.46*	0.30	0.77***
	(0.77)	(0.67)	(0.66)	(0.56)	(0.49)	(0.47)	(0.24)	(0.26)	(0.23)
Constant	−101.85	−60.54	−106.83*	−90.57*	−43.47	−87.51*	−58.39***	11.32	−33.49
	(66.54)	(65.93)	(63.35)	(48.45)	(48.80)	(44.59)	(20.94)	(25.56)	(22.46)

Notes. Sample 121 female authors; 104 male authors. Sample restricted to matched authors. See Section 3.4.3 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.

D.7 Section 3.4.3, list of matched pairs. Table D.15 displays the names of the economists in each matched pair.

Table D.15: Matched pairs

Matched pairs		Matched pairs	
Female	Male	Female	Male
Abraham, Katharine G.	Rhee, Changyong	La Ferrara, Eliana	Bó, Pedro Dal
Admati, Anat R.	Rhee, Changyong	Landes, Elisabeth M.	Friedman, David
Amiti, Mary	Salvanes, Kjell G.	Levy, Gilat	Strumpf, Koleman S.
Anderson, Siwan	Michalopoulos, Stelios	Lewis, Karen K.	Gale, William G.
Ashraf, Nava	Seshadri, Ananth	Li, Wei	Strumpf, Koleman S.
Athey, Susan	Kmenta, Jan	Lleras-Muney, Adriana	von Wachter, Till
Baicker, Katherine	Van Reenen, John	Løken, Katrine Vellesen	Gertler, Paul J.
Bailey, Martha J.	Doepke, Matthias	Madrian, Brigitte C.	Hassler, John
Bandiera, Oriana	Staiger, Douglas O.	Maestas, Nicole	Naidu, Suresh
Barwick, Panle Jia	Goyal, Sanjeev	Malmendier, Ulrike	Rubinfeld, Daniel L.
Baxter, Marianne	Fershtman, Chaim	Matzkin, Rosa L.	Mullainathan, Sendhil
Bedard, Kelly	Mahajan, Aprajit	McConnell, Sheena	Oyer, Paul
Bertrand, Marianne	Ray, Edward John	McGrattan, Ellen R.	Flinn, Christopher J.
Black, Sandra E.	Cahuc, Pierre	Meyer, Margaret A.	Gennaioli, Nicola
Blank, Rebecca M.	Naidu, Suresh	Molinari, Francesca	Vermeulen, Frederic
Boustan, Leah Platt	Pope, Devin G.	Moser, Petra	Dahl, Gordon B.
Brown, Jennifer	Gale, William G.	Nakamura, Emi	Snowberg, Erik
Busse, Meghan R.	La Porta, Rafael	Ng, Serena	Renault, Eric
Case, Anne C.	Thomson, William	Niederle, Muriel	Board, Simon
Casella, Alessandra	Mendelsohn, Robert	Oster, Emily	Kremer, Michael
Chen, Xiaohong	Wilson, John Douglas	Pande, Rohini	Kane, Thomas J.
Chen, Yan	Irwin, Douglas A.	Paxson, Christina H.	Pauzner, Ady
Chevalier, Judith A.	Eliaz, Kfir	Perrigne, Isabelle	Rhee, Changyong
Chichilnisky, Graciela	Hubbard, Thomas N.	Piazzesi, Monika	Kahn, James A.
Correia, Isabel	Bohn, Henning	Qian, Nancy	Kahn, Matthew E.
Costa, Dora L.	Dorn, David	Quinzii, Martine	Williams, Steven R.
Cropper, Maureen L.	Strahan, Philip E.	Ramey, Valerie A.	Evans, Paul
Currie, Janet	Kosfeld, Michael	Reinganum, Jennifer F.	Manski, Charles F.
Dafny, Leemore S.	Xu, Daniel Yi	Reinhart, Carmen M.	Lefgren, Lars
De Nardi, Mariacristina	Kosfeld, Michael	Rey, Hélène	Waugh, Michael E.
Demange, Gabrielle	Roemer, John E.	Romer, Christina D.	Cooley, Thomas F.
Duflo, Esther	Bettinger, Eric P.	Rose, Nancy L.	Snowberg, Erik
Dupas, Pascaline	Kremer, Michael	Rose-Ackerman, Susan	Mookherjee, Dilip
Dynan, Karen E.	Wiggins, Steven N.	Rosenblat, Tanya S.	Guryan, Jonathan
Eberly, Janice C.	Einav, Liran	Rouse, Cecilia Elena	Black, Dan A.
Eckel, Catherine C.	Grinblatt, Mark S.	Sapienza, Paola	Verdier, Thierry
Edlund, Lena	van Wijnbergen, Sweder	Schennach, Susanne M.	Burnside, Craig
Eyigungor, Burcu	McClellan, Mark B.	Schmitt-Grohé, Stephanie	Woodford, Michael
Fan, Yanqin	Matsusaka, John G.	Schwartz, Nancy L.	Shimer, Robert
Fernández, Raquel	Svensson, Jakob	Shannon, Chris	Williams, Steven R.
Field, Erica	Kremer, Michael	Shaw, Kathryn L.	Gould, Eric D.
Finkelstein, Amy	Sacerdote, Bruce I.	Spier, Kathryn E.	Chay, Kenneth Y.
Flavin, Marjorie A.	Eyster, Erik	Stokey, Nancy L.	Hynes, J. Allan
Forges, Françoise	Christensen, Laurits R.	Tenreyro, Silvana	Skinner, Jonathan
Fortin, Nicole M.	Sacerdote, Bruce I.	Tertilt, Michèle	Hyslop, Dean R.
Freund, Caroline	Bernard, Andrew B.	Tesar, Linda L.	Meyer, Bruce D.
Fuchs-Schündeln, Nicola	Marcet, Albert	Thomas, Julia K.	Rhee, Changyong
Garfinkel, Michelle R.	Finan, Frederico	Todd, Petra E.	Sanders, Seth G.
Goldberg, Pinelopi Koujianou	Burstein, Ariel Tomás	Vissing-Jørgensen, Annette	MacLeod, W. Bentley
Goldin, Claudia D.	Boldrin, Michele	Voena, Alessandra	Donohue, John J. (III)
Gopinath, Gita	Chetty, Raj	Washington, Ebonya L.	Oyer, Paul

Table D.15 (continued)

Matched pairs		Matched pairs	
Female	Male	Female	Male
Griffith, Rachel	Oreopoulos, Philip	White, Lucy	Strumpf, Koleman S.
Guerrieri, Veronica	Hillman, Arye L.	Whited, Toni M.	Jansson, Michael
Hanna, Rema	Möbius, Markus M.	Williams, Heidi L.	Rockoff, Jonah E.
Hastings, Justine S.	Ferrie, Joseph P.	Wooders, Myrna Holtz	Isaac, R. Mark
Ho, Katherine	Nunn, Nathan	Yariv, Leeat	Finan, Frederico
Hoxby, Caroline Minter	Goldfarb, Avi	Yellen, Janet L.	Rogerson, Richard
Jayachandran, Seema	Dahl, Gordon B.	Zeiler, Kathryn	McAdams, David
Kowalski, Amanda E.	Munshi, Kaivan	Zhuravskaya, Ekaterina	Knittel, Christopher R.
Kranton, Rachel E.	Rockoff, Jonah E.	İmrohoroğlu, Ayşe	Kircher, Philipp
Kuziemko, Ilyana	Graham, Bryan S.		

Notes. Table lists the names of the matched pairs from Section 3.4.3. In each panel, female members are listed first; male members second. Matches were made using a probit model with replacement. See Section 3.4.3 for details on the matching process.

D.8 Table X, male effects, Equation (12) and Condition 1. Table D.16 estimates D_{ik} with Equation (12). Table X estimates D_{ik} with a rough attempt to control for acceptance rates—it requires $T_i \leq T_k$ or $T_k \leq T_i$ before categorising matched pairs as discrimination against i or k, respectively. Conclusions from both tables are are similar to those presented in Section 3.4.3. Table D.18 shows $\widehat{R}_{k3}$ for men in the matched sample. Grade-level effects (Flesch-Kincaid, Gunning Fog, SMOG and Dale-Chall) have been multiplied by negative one (Section 2.1).

TABLE D.16: D_{ik}, Equation (12)

	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	10.51	8.55	58	−9.06	8.19	21	3.87***	3.20***
							(1.09)	(1.16)
Flesch Kincaid	2.04	1.80	61	−1.14	1.45	25	0.85***	0.68***
							(0.22)	(0.24)
Gunning Fog	3.05	2.16	62	−1.85	1.87	17	1.40***	1.20***
							(0.27)	(0.29)
SMOG	2.07	1.60	63	−1.53	1.50	16	0.91***	0.77***
							(0.19)	(0.21)
Dale-Chall	0.90	0.68	48	−0.53	0.49	23	0.26***	0.20**
							(0.08)	(0.09)

Notes. Sample 121 matched pairs (104 and 121 distinct men and women, respectively). First and second panels display conditional means, standard deviations and observation counts of $\underline{D}_{ik}$ (Equation (12)) 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.

TABLE D.17: 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	15.38	10.43	40	−11.65	11.50	12	5.49***	4.65***
							(1.59)	(1.69)
Flesch Kincaid	2.98	2.00	40	−2.03	2.34	12	1.15***	0.92**
							(0.34)	(0.37)
Gunning Fog	4.24	2.69	44	−3.57	2.86	9	1.76***	1.51***
							(0.42)	(0.45)
SMOG	3.03	1.92	45	−2.90	2.18	8	1.26***	1.09***
							(0.30)	(0.32)
Dale-Chall	1.80	1.35	30	−0.91	0.56	15	0.45***	0.38**
							(0.16)	(0.17)

Notes. Sample 121 matched pairs (104 and 121 distinct men and women, respectively). First and second panels display conditional means, standard deviations and observation counts of $\underline{D}_{ik}$ (Equation (11)) from subpopulations of matched pairs in which the woman or man, respectively, satisfies Conditions 1–3. 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.

TABLE D.18: Mean $\widehat{R}_{k3}$ (men)

	Flesch Reading Ease	Flesch-Kincaid	Gunning Fog	SMOG	Dale-Chall
$\widehat{R}_{k3}$ (men)	36.83	−13.95	−18.32	−16.00	−11.54
	(1.114)	(0.243)	(0.282)	(0.198)	(0.100)

Notes. Sample 121 matched pairs (104 and 121 distinct men and women, respectively). Figures correspond to the $t = 3$ reconstructed readability scores for men. $\widehat{R}_{i3}$ weighted by frequency observations are used in a match; degrees-of-freedom corrected standard errors in parentheses.

D.9 Table XI, alternative thresholds for mother$_j$. Table D.19 repeats the regression presented in Table XI 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 D.19: Table XI, alternative thresholds for mother$_j$

	Age < 3	Age < 4	Age < 5	Age < 10	Age < 18
Female ratio	5.653***	6.341***	6.654***	6.562***	6.335***
	(2.102)	(2.097)	(2.150)	(2.175)	(2.225)
Mother	−3.673	−11.068***	−10.934***	−8.914**	−5.550
	(2.327)	(3.599)	(3.212)	(3.495)	(3.399)
Birth	1.317	7.999*	7.579*	5.651	2.518
	(3.784)	(4.464)	(4.167)	(4.402)	(4.126)
Max. t_j	−0.163**	−0.165**	−0.163**	−0.163**	−0.162**
	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)
No. pages	0.180***	0.178***	0.178***	0.178***	0.179***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
N	1.005**	0.979**	0.970**	0.968**	0.975**
	(0.443)	(0.443)	(0.443)	(0.444)	(0.445)
Order	0.221**	0.220**	0.220**	0.218**	0.219**
	(0.089)	(0.089)	(0.089)	(0.089)	(0.089)
No. citations	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	37.732***	37.866***	37.892***	37.866***	37.781***
	(2.049)	(2.054)	(2.057)	(2.059)	(2.047)
Editor effects	✓	✓	✓	✓	✓
Year effects	✓	✓	✓	✓	✓
Institution effects	✓	✓	✓	✓	✓

Notes. Sample 2,626 articles. Coefficients from OLS estimation of Equation (13) 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 XI. Standard errors clustered by year in parentheses. ***, ** and * statistically significant at 1%, 5% and 10%, respectively.

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.

Ali, O. et al. (2010). "Automating News Content Analysis: An Application to Gender Bias and Readability". *Workshop on Applications of Pattern Analysis* 11, pp. 36–43.

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

Antecol, H., K. Bedard, and J. Stearns (2016). "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" IZA Discussion Paper Series, No. 9904.

Anzia, S. F. and C. R. Berry (2011). "The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55 (3), pp. 478–493.

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

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), p. 277.

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.

Arrow, K. (1973). "The Theory of Discrimination". In: *Discrimination in Labor Markets*. Ed. by O. Ashenfelter and A. Rees. Princeton, New Jersey: Princeton University Press. Chap. 1, pp. 16–195.

Artz, B., A. H. Goodall, and A. J. Oswald (2016). "Do Women Ask ?" IZA Discussion Paper Series, No. 10183.

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.

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

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

Baker, K. J. (2015). "Should Academic Conferences Have Codes of Conduct?" *Chronicle Vitae*. https://chroniclevitae.com/news/1182-should-academic-conferences-have-codes-of-conduct. Accessed: 2016-10-04.

Bandiera, O. (2016). *The Gender and Ethnicity Earnings Gap at LSE*. Tech. rep. September. London School of Economics.

Bazargan, M. and V. S. Guzhva (2011). "Impact of Gender, Age and Experience of Pilots on General Aviation Accidents". *Accident Analysis and Prevention* 43 (3), pp. 962–970.

Becker, G. S. (1957). *The Economics of Discrimination*. 2nd ed. Chicago, Illinois: University of Chicago Press.

Begeny, J. C. and D. J. Greene (2014). "Can Readability Formulas Be Used to Successfully Gauge Difficulty of Reading Materials?" *Psychology in the Schools* 51 (2), pp. 198–215.

Benedetti, T. J. et al. (2004). "The Productivity of Washington State's Obstetrician–Gynecologist Workforce: Does Gender Make a Difference?" *Obstetrics and Gynecology* 103 (3), pp. 499–505.

Benoit, K., K. Munger, and A. Spirling (2017). "Measuring and Explaining Political Sophistication through Textual Complexity". Mimeo.
Berk, J. B., C. R. Harvey, and D. Hirshleifer (2017). "How to Write an Effective Referee Report and Improve the Scientific Review Process". *Journal of Economic Perspectives* 31 (1), pp. 231–244.
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.
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.
Biddle, C. and J. Aker (1996). "How Does the Peer Review Process Influence AANA Journal Article Readability?" *AANA Journal* 64 (1), pp. 65–68.
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.
Blau, F. D. and L. M. Kahn (2016). "The Gender Wage Gap: Extent, Trends, and Explanations". NBER Working Paper Series, No. 21913.
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.
Blundell, R. and S. Bond (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". *Journal of Econometrics* 87 (1), pp. 115–143.
Bordalo, P. et al. (2016). "Stereotypes". *Quarterly Journal of Economics* 131 (4), pp. 1753–1794.
Boring, A. (2017). "Gender Biases in Student Evaluations of Teaching". *Journal of Public Economics* 145 (Supplement C), pp. 27–41.
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.
Bransch, F. and M. Kvasnicka (2017). "Male Gatekeepers Gender Bias in the Publishing Process?" IZA Discussion Paper Series, No. 11089.
Bright, L. K. (2017). "Decision Theoretic Model of the Productivity Gap". *Erkenntnis* 82 (2), pp. 421–442.
Budden, A. E. et al. (2008a). "Double-blind Review Favours Increased Representation of Female Authors". *Trends in Ecology and Evolution* 23 (1), pp. 4–6.
Budden, A. E. et al. (2008b). "Response to Webb et al.: Double-blind Review: Accept with Minor Revisions". *Trends in Ecology and Evolution* 23 (7), pp. 353–354.
Budden, A. E. et al. (2008c). "Response to Whittaker: Challenges in Testing for Gender Bias". *Trends in Ecology and Evolution* 23 (9), pp. 480–481.
Canadian Institute for Health Information (2005). *Canada's Health Care Providers*. https://secure.cihi.ca/free_products/HCP_Chartbook05_e.pdf. Accessed: 2017-01-10.
Card, D. and S. DellaVigna (2013). "Nine Facts about Top Journals in Economics". *Journal of Economic Literature* 51 (1), pp. 144–161.
— (2017). "What do Editors Maximize? Evidence from Four Leading Economics Journals". NBER Working Paper Series, No. 23282.
Ceci, S. J. et al. (2014). "Women in Academic Science: A Changing Landscape". *Psychological Science in the Public Interest* 15 (3), pp. 75–141.

Chall, J. S. and E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.
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.
Chung, J. and G. S. Monroe (2001). "A Research Note on the Effects of Gender and Task Complexity on an Audit Judgment". *Behavioral Research in Accounting* 13 (1), pp. 111–125.
Clain, S. H. and K. Leppel (2017). "Patterns in Economics Journal Acceptances and Rejections". *American Economist* (forthcoming).
Coate, S. and G. C. Loury (1993). "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5), pp. 1220–1240.
Correll, S. and C. Simard (2016). "Research: 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.
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.
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.
Craig, A. and R. Fryer (2017). "Complementary Bias: A Model of Two-Sided Statistical Discrimination". NBER Working Paper Series, No. 23811.
Dale, E. and J. S. Chall (1948). "A Formula for Predicting Readability". *Educational Research Bulletin* 27 (1), pp. 11–20.
DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information.
Ecklund, E. H. and A. E. Lincoln (2011). "Scientists Want More Children". *PLoS ONE* 6 (8), pp. 1–4.
Ellison, G. (2002). "The Slowdown of the Economics Publishing Process". *Journal of Political Economy* 110 (5), pp. 947–993.
Faigley, L. and S. P. Witte (1981). "Analyzing Revision". *College Composition and Communication* 32 (4), pp. 400–414.
Family, H., M. Weiss, and J. Sutton (2013). *The Effects of Mental Workload on Community Pharmacists' Ability to Detect Dispensing Errors*. Tech. rep. Pharmacy Research UK.
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.
Feingold, A. (1994). "Gender Differences in Personality: A Meta-analysis". *Psychological Bulletin* 116 (3), pp. 429–456.
Firth-Cozens, J. (2008). "Doctors with Difficulties: Why So Few Women?" *Postgraduate Medical Journal* 84 (992), pp. 318–320.
Foschi, M. (1996). "Double Standards in the Evaluation of Men and Women". *Social Psychology Quarterly* 59 (3), pp. 237–254.
Francis, D. J. et al. (2008). "Form Effects on the Estimation of Students' Oral Reading Fluency Using DIBELS". *Journal of School Psychology* 46 (3), pp. 315–342.
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.

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.
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.
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.
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.
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.
Goldberg, P. (1968). "Are Women Prejudiced against Women?" *Trans-action* 5 (5), pp. 28–30.
Goldberg, P. K. (2015). "Report of the Editor: American Economic Review". *American Economic Review* 105 (5), pp. 698–710.
Goldin, C. (2014). "A Grand Gender Convergence: Its Last Chapter". *American Economic Review* 104 (4), pp. 1091–1119.
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.
Goldin, C. and C. Rouse (2000). "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians". *American Economic Review* 90 (4), pp. 715–741.
Gordon, M. B. et al. (2009). "Gender Differences in Research Grant Applications for Pediatric Residents". *Pediatrics* 124 (2), e355–61.
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.
Hamermesh, D. S. (1994). "Facts and Myths about Refereeing". *Journal of Economic Perspectives* 8 (1), pp. 153–163.
Hartley, J., J. W. Pennebaker, and C. Fox (2003a). "Abstracts, Introductions and Discussions: How Far Do They Differ in Style?" *Scientometrics* 57 (3), pp. 389–398.
— (2003b). "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.
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.
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.
Hayden, J. D. (2008). "Readability in the British Journal of Surgery". *British Journal of Surgery* 95 (1), pp. 119–124.
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.
Hengel, E. (2015). "Two Essays on Bankruptcy and One Essay on Gender Differences in Academic Publishing". PhD thesis. University of Cambridge.
— (2016). "Publishing while Female: Gender Differences in Peer Review Scrutiny". Mimeo.

Hintze, J. M. and T. J. Christ (2004). "An Examination of Variability as a Function of Passage Variance in CBM Progress Monitoring". *School Psychology Review* 33 (2), pp. 204–217.
Ittonen, K., E. Vähämaa, and S. Vähämaa (2013). "Female Auditors and Accruals Quality". *Accounting Horizons* 27 (2), pp. 205–228.
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.
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.
Journal of Trauma and Acute Care Editorial Board (2015). *Journal of Trauma and Acute Care Surgery*. http://journals.lww.com/jtrauma/Documents/Editorial%20Board%20Reports/2015%200909%20AAST%20Ed%20Board%20print%20report.pdf. Accessed: 2016-10-04.
Kimble, J. (1994). "Answering the Critics of Plain Language". *Scribes Journal of Legal Writing* 51 (1994-1995), pp. 51–85.
King, D. W., C. Tenopir, and M. Clarke (2006). "Measuring Total Reading of Journal Articles". *D-Lib Magazine* 12 (10), pp. 1082–9873.
Klos, D. M. (2014). *The Status of Women in the U.S. Media 2013*. Tech. rep. Women's Media Center.
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.
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.
Kugler, A., C. Tinsley, and O. Ukhaneva (2017). "Choice of Majors: Are Women Really Different from Men?" NBER Working Paper Series, No. 23735.
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.
Lee, C. J. (2016). "Revisiting Current Causes of Women's Underrepresentation in Science". In: *Metaphysics and Epistemology*. Ed. by M. Brownstein and J. Saul. Vol. 1. Oxford: Oxford University Press. Chap. 2.5, pp. 265–283.
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.
Liang, F. M. (1983). "Word Hy-phen-a-tion by Com-put-er". PhD thesis. Stanford University.
Long, L. N. and W. F. Christensen (2011). "Does the Readability of Your Brief Affect Your Chance of Winning an Appeal?" *Journal of Appellate Practice and Process* 12 (1), pp. 1–14.
Loughran, T. and B. McDonald (2014). "Measuring Readability in Financial Disclosures". *Journal of Finance* 69 (4), pp. 1643–1671.
McFadden, K. L. (1996). "Comparing Pilot-error Accident Rates of Male and Female Airline Pilots". *Omega* 24 (4), pp. 443–450.
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.

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.
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.
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.
Niskanen, J. et al. (2011). "Auditor Gender and Corporate Earnings Management Behavior in Private Finnish Firms". *Managerial Auditing Journal* 26 (9), pp. 778–793.
O'Donnell, E. and E. N. Johnson (2001). "The Effects of Auditor Gender and Task Complexity on Information Processing Efficiency". *International Journal of Auditing* 5 (2), pp. 91–105.
Paludi, M. A. and W. D. Bauer (1983). "Goldberg Revisited: What's in an Author's Name". *Sex Roles* 9 (3), pp. 387–390.
Parsons, C. A. et al. (2011). "Strike Three: Discrimination, Incentives, and Evaluation". *American Economic Review* 101 (4), pp. 1410–1435.
Payne, B. K. and D. Dabney (1997). "Prescription Fraud: Characteristics, Consequences, and Influences". *Journal of Drug Issues* 27 (4), pp. 807–820.
Pertold-Gebicka, B., F. Pertold, and N. D. Gupta (2016). "Employment Adjustments around Childbirth". IZA Discussion Paper Series, No. 9685.
Phelps, E. S. (1972). "The Statistical Theory of Racism and Sexism". *American Economic Review* 62 (4), pp. 659–661.
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.
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.
Roberts, T.-A. and S. Nolen-Hoeksema (1989). "Sex Differences in Reactions to Evaluative Feedback". *Sex Roles* 21 (11-12), pp. 725–747.
— (1994). "Gender Comparisons in Responsiveness to Others' Evaluations in Achievement Settings". *Psychology of Women Quarterly* 18, pp. 221–240.
Romero, J. (2013). "Where Are the Women?" *Econ Focus* 7 (2), p. 12.
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.
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 (February), pp. 811–825.
Sarsons, H. (2016). "Gender Differences in Recognition for Group Work". Mimeo.
— (2017). "Recognition for Group Work: Gender Differences in Academia". *American Economic Review* 107 (5), pp. 141–145.
Schafheutle, E. I., E. M. Seston, and K. Hassell (2011). "Factors Influencing Pharmacist Performance: A Review of the Peer-Reviewed Literature". *Health Policy* 102 (2–3), pp. 178–192.
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.

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.
Sheltzer, J. M. and J. C. Smith (2014). "Elite Male Faculty in the Life Sciences Employ Fewer Women". *Proceedings of the National Academy of Sciences* 111 (28), pp. 10107–10112.
Sirico, L. J. (2007). "Readability Studies: How Technocentrism Can Compromise Research and Legal Determinations". *Quinnipiac Law Review* 26 (1), pp. 147–172.
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.
Stempel, G. H. (1981). "Readability of Six Kinds of Content in Newspapers". *Newspaper Research Journal* 3 (1), pp. 32–37.
Szeinbach, S. et al. (2007). "Dispensing Errors in Community Pharmacy: Perceived Influence of Sociotechnical Factors". *International Journal for Quality in Health Care* 19 (4), pp. 203–209.
Thörnqvist, T. (2015). "Sophistication, News and Individual Investor Trading". Mimeo.
Torgler, B. and M. Piatti (2013). *A Century of American Economic Review*. New York, New York: Palgrave Macmillan.
Tsugawa, Y. et al. (2016). "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs. Female Physicians". *JAMA Internal Medicine* 02138, pp. 1–8.
Tullett, J., P. Rutter, and D. Brown (2003). "A Longitudinal Study of United Kingdom Pharmacists' Misdemeanours—Trials, Tribulations and Trends". *Pharmacy World & Science* 25 (2), pp. 43–51.
Turnbull, G. K. and J. Dombrow (2007). "Individual Agents, Firms, and the Real Estate Brokerage Process". *Journal of Real Estate Finance and Economics* 35 (1), pp. 57–76.
Vail, G. J. and L. G. Ekman (1986). "Pilot-error Accidents: Male vs. Female". *Applied Ergonomics* 17 (4), pp. 297–303.
Van Rooyen, S., T. Delamothe, and S. J. W. Evans (2010). "Effect on Peer Review of Telling Reviewers that Their Signed Reviews Might Be Posted on the Web: Randomised Controlled Trial". *British Medical Journal* 341 (c5729).
Van Rooyen, S. et al. (1999). "Effect of Open Peer Review on Quality of Reviews and on Reviewers' Recommendations: A Randomised Trial". *British Medical Journal* 318 (7175), pp. 23–27.
Volden, C., A. E. Wiseman, and D. E. Wittmer (2013). "When Are Women More Effective Lawmakers Than Men?" *American Journal of Political Science* 57 (2), pp. 326–341.
Voyer, D. and S. D. Voyer (2014). "Gender Differences in Scholastic Achievement: A Meta-Analysis". *Psychological Bulletin* 140 (4), pp. 1174–1204.
Walsh, E. et al. (2000). "Open Peer Review: A Randomised Controlled Trial". *British Journal of Psychiatry* 176 (1), pp. 47–51.
Walton, R. O. and P. M. Politano (2016). "Characteristics of General Aviation Accidents Involving Male and Female Pilots". *Aviation Psychology and Applied Human Factors* 6 (1), pp. 39–44.
Webb, T. J., B. O'Hara, and R. P. Freckleton (2008). "Does Double-blind Review Favor Female Authors?" *Trends in Ecology and Evolution* 6 (7), pp. 351–353.
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 (1-11).

Weisshaar, K. (2017). "Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia". *Social Forces* (November), pp. 1–31.
Whittaker, R. J. (2008). "Journal Review and Gender Equality: A Critical Comment on Budden et al." *Trends in Ecology and Evolution* 23 (9), pp. 478–479.
Williams, J. C., K. W. Phillips, and E. V. Hall (2015). *Double Jeopardy? Gender Bias against Women of Color in Science*. Tech. rep. University of California, Hastings College of the Law.
Wu, A. H. (2017). "Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum". Mimeo.
Xie, Y. and K. A. Shauman (2005). *Women in Science: Career Processes and Outcomes*. Cambridge, Massachussets: Harvard University Press.